HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

BACHELORARBEIT

Herr
Martin Garbe

Erkennung von Ähnlichkeitsbeziehungen wohldefinierter Muster mit Hilfe regulärer Ausdrücke

Mittweida, 2012

Fakultät Mathematik, Naturwissenschaften,
Informatik

BACHELORARBEIT

Erkennung von Ähnlichkeitsbeziehungen wohldefinierter Muster in Membranproteinen mit Hilfe regulärer Ausdrücke

Autor:
Herr

Martin Garbe

Studiengang:
Biotechnologie/Bioinformatik

Seminargruppe:
BI09w1-B

Erstprüfer:
M. Sc. Steffen Grunert

Zweitprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Einreichung:
Mittweida, 23.08.2012

Verteidigung:
Mittweida,2012

Faculty mathematics, natural sciences,
informatics

BACHELOR THESIS

Recognition of similarity relations of well-defined pattern using regular expressions

author:
Mr.

Martin Garbe

course of studies:
Biotechnology/Bioinformatics

seminar group:
BI09w1-B

first examiner:
M. Sc. Steffen Grunert

second examiner:
Prof. Dr. rer. nat. Dirk Labudde

submission:
Mittweida, 23.08.2012

defence/ evaluation:
Mittweida,2012

Bibliographische Beschreibung:

Garbe, Martin:
Erkennung von Ähnlichkeitsbeziehungen wohl definierter Muster in Membranproteinen mit Hilfe regulärer Ausdrücke. - 2012. – 64 S.
Mittweida, Hochschule Mittweida, Fakultät Mathematik, Naturwissenschaften, Informatik, Bachelorarbeit, 2012

Referat:

In der Bachelorarbeit ,,Erkennung von Ähnlichkeitsbeziehungen wohl definierter Muster in Membranproteinen mit Hilfe regulärer Ausdrücke" wurde 32 Pfamfamilien, in Bezug auf Motive, untersucht.Als Grundlage dienten hierfür 50 Sequenzmuster von Gerstein et al. [11].
Die Proteinsequenzen der verwendeten 32 Proteinfamilien, von deren Domänen keine Funktion bekannt ist, wurden auf Beinhalten dieser Muster hin untersucht. Im nächsten Schritt wurde untersucht, welche dieser 50 Muster zusammen in einer Proteinsequenz auftreten, um so die Co - Co- Occurence festzustellen. Anhand der somit erhaltenen häufigsten Paarungen im transmembranen, nichttransmembranen und Transition - Bereich und mit Hilfe von bereits beschriebenen Mustern und Pattern von biologischen Datenbanken, wurde versucht, bestimmten Musterpaarungen aus der Arbeit von Gerstein et al. eine Funktion zuzuordnen bzw. eine Erklärung für ihr Auftreten in den Proteinen zu finden.

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich bei der Anfertigung
dieser Bachelor-Thesis unterstützt haben.

Ein besonderer Dank gilt meinen beiden Betreuern, Prof. Dr. rer. nat. Dirk Labudde, welcher mich immer wieder durch seine Anregungen und Erklärungen aus einer anderen Perspektive als der meinigen unterstützt hat und mir half, die Struktur dieser Arbeit zu festigen, sowie
M. Sc. Steffen Grunert, der mir bei der Implementierung von Programmen und dem Verständis der zu bewältigenden Aufgaben eine wichtige Hilfe war.

Des Weiteren möchte ich mich bei Jenny Bleeck bedanken, die mich bei meiner Arbeit bei Analysen unterstützt hat.
Nicht zuletzt möchte ich meiner Familie und meinen Freunden danken, die mich stets moralisch unterstützt haben und im Besonderen meiner Freundin, welche dafür Sorge getragen hat, dass ich mich stets auf meine Arbeit konzentrieren konnte.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

Abkürzungsverzeichnis

A	Alanin
AS	Aminosäure
BLAST	Basic Local Alignment Search Tool
C	Cystein
CD – HIT	Cluster Database at High Identity with Tolerance
D	Asparaginsäure
E	Glutaminsäure
ELM	Eukaryotic Linear Motif
F	Phenylalanin
G	Glycin
GG4	G-X-X-X-G Sequenzmotiv
H	Histidin
HF	Häufiigkeit
HMM	Hidden Markov Model
I	Isoleucin
K	Lysin
L	Leucin
M	Methionin
N	Asparagin
nTM	nichttransmembraner Bereich
P	Prolin
Q	Glutamin
R	Arginin
S	Serin
T	Threonin
TM	transmembraner Bereich
TMHMM	Transmembrane Hidden Markov Model
V	Valin
W	Tryptophan
Y	Tyrosin
X	beliebige Aminosäure

1 Motivation

Bei Sequenzmotiven handelt es sich um lokal konservierte Regionen einer Sequenz oder auch um Muster, die in verschiedenen Sequenzen auftreten. Sie dienen vor allem zur Erkennnung von strukturellen oder funktionellen Eigenschaften des Proteins und können auch dazu genutzt werden, die Zugehörigkeit von Proteinen zu einer Sequenzfamilie zu definieren. Membranproteine sind an einem Großteil der biologischen Prozesse beteiligt, weshalb versucht wird, ihre genaue Funktion bei diesen Prozessen aufzuklären und den Grund dafür in ihren Proteinstrukturen zu suchen. Verschiedene Muster sorgen in Verbindung mit weiteren für unterschiedliche Bindungseigenschaften innerhalb eines Proteins und bestimmen somit zum einen die Tertiärstruktur des Proteins, zum anderen aber auch die jeweilige Funktion. Des weiteren haben die physikochemischen Eigenschaften, welche durch ihre enthaltenen Aminosäuren und deren Reste bestimmt werden, der einzelnen Motive auch Einfluss auf Motive in anderen Bereichen des Proteins. Sollten immer wieder die selben in bestimmten Regionen von Proteinen vorkommen, könnte dies auch für eine bestimme Funktion sprechen beziehungsweise könnten diese Muster dafür verantwortlich sein, welche Muster darauffolgend auftreten, da sie durch ihre physikochemischen Eigenschaften eben den weiteren Verlauf der Proteinsequenz mit beeinflussen.

2 Grundlagen

2.1 Aufbau und Funktion von Proteinen

Jedes Lebewesen besteht aus einer Vielzahl von Zellen und zu deren Grundbausteinen gehören unter anderem die Proteine. Die Gesamtheit aller Proteine im Lebewesen, Gewebe oder auch nur in einer Zelle betrachtet wird als Proteom bezeichnet.

Es handelt sich um organische Makromoleküle, welche eine Molekülmasse zwischen 1000 und 1000000 Dalton besitzen. Diese Makromoleküle wiederum sind aus den 20 kanonischen Aminosäuren aufgebaut, welche in Tabelle 1 dargestellt sind [1].

Tabelle 1 Übersicht über die 20 kanonischen Aminosäuren

Tabellle der 20 kanonischen Aminosäuren in Ein- und Drei-Letter-Code.

Aminosäure	Ein-Letter-Code	Drei-Letter-Code
Alanin	A	Ala
Cystein	C	Cys
Asparaginsäure	D	Asp
Glutaminsäure	E	Glu
Phenylalanin	F	Phe
Glycin	G	Gly
Histidin	H	His
Isoleucin	I	Ile
Lysin	K	Lys
Leicin	L	Leu
Methionin	M	Met
Asparagin	N	Asn
Prolin	P	Pro
Glutamin	Q	Gln
Arginin	R	Arg
Serin	S	Ser
Threonin	T	Thr
Valin	V	Val
Tryptophan	W	Trp
Tyrosin	Y	Tyr

Jede dieser Aminosäuren besitzt einen charakteristischen Grundaufbau aus einer Aminogruppe und einer Carboxylgruppe am α-Kohlenstoffatom, erst der jeweilige Rest bestimmt die physikochemischen Eigenschaften der Aminosäure (siehe Abbildung 1).

Um ein Makromolekül zu bilden, wird eine Peptidbindung zwischen der Amino- und der Carboxylgruppe ausgebildet, haben sich auf diese Weise mindestens 100 Aminosäuren aneinandergelagert, so spricht man von einem Protein.

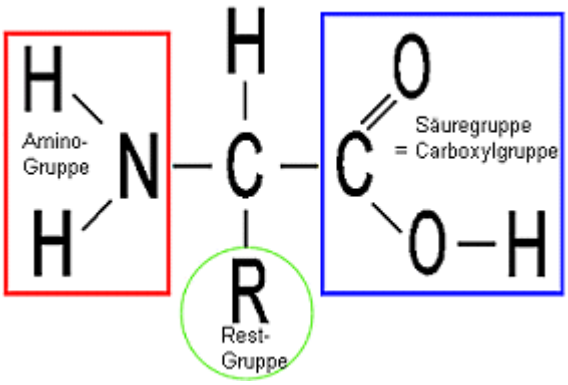

Abbildung 1 Grundaufbau einer Aminosäure

Abbildung des Grundaufbaus einer Aminosäure, bestehend aus einer Aminogruppe (rot markiert), einer Carboxylgruppe (blau markiert) und der Restgruppe (grün markiert).

Durch die Restgruppe und die daraus resultierenden Eigenschaften der einzelnen Aminosäuren, können diese auch in verschiedene Gruppen eingeteilt werden. Aminosäuren können einen hydrophoben Charakter besitzen, was bedeutet, dass sie sich nicht in Wasser lösen und es sogar eher abstoßen, weshalb hydrophobe Aminosäuren vor allem im Inneren der Transmembran zu finden sind, im Gegensatz zu den hydrophilen Aminosäuren, welche dem Wasser zugewandt sind und die hydrophoben Aminosäuren in flüssigem Milieu umlagern. Andere Eigenschaften sind z. Bsp. Polarität und Größe, da eine einzelne Aminosäure verschiedene Eigenschaften gleichzeitig besitzt, ist die anschauliche Darstellung der Einteilung in Gruppen durch sogenannte Venn – Diagramme realisiert worden, in Abbildung 2 ist eines für die 20 kanonischen Aminosäuren abgebildet.

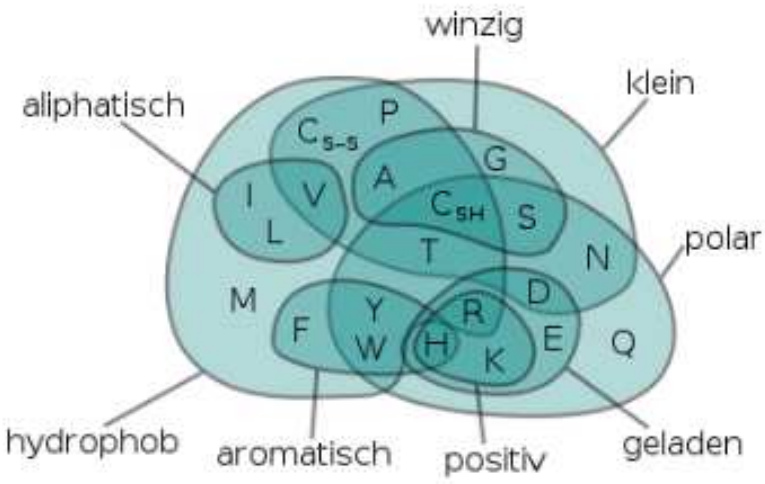

Abbildung 2 Venn – Diagramm

In diesem Venn – Diagramm sind die Aminosäuren nach ihren jeweiligen Eigenschaften gruppiert.

Durch das Zusammenwirken der Aminosäuren wird auch das Protein in seinen Eigenschaften beeinflusst. Hauptsächlich bestimmt werden die Eigenschaften der Proteine durch ihre charakteristische dreidimensionale Struktur, welche wiederum durch die Sekundär- und Primärstruktur beeinflusst wird.

Die Primärstruktur ist die jeweilige Aminosäuresequenz des Proteins, sprich die Anordnung hintereinander in einer Kette. Bei der Sekundärstruktur handelt es sich um den Aufbau des ,,Gerüstes" des Proteins aus Sekundärstrukturelementen wie Helices, Faltblattstrukturen oder Loops. Als Tertiärstruktur bezeichnet man die dreidimensionale Struktur, also die Faltung der Sekundärstrukturelemente, wie in Abbildung 3 zu erkennen ist und die Anordnung der Seitenketten im Raum. Größere Proteine falten sich meist in sogenannten Domänen, welche aus 100 – 200 Aminosäuren bestehen und einen Durchmesser von etwa 2,5 nm besitzen.

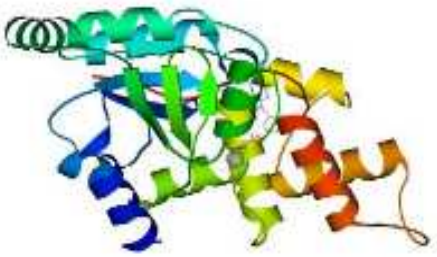

Abbildung 3 Tertiärstruktur eines Proteins

Ein Beispiel für die Tertiärstruktur eines Proteins, zu erkennen ist die Anordnung der Sekundärstrukturelemente im Raum.

Proteine sind für eine Vielzahl von Stofftransporten verantwortlich und lassen sich zudem aufgrund ihrer unterschiedlichen Struktur und Eigenschaften in verschiedene Gruppen einteilen. Membranproteine stellen solch eine Gruppe dar, sie verhalten sich anders als lösliche Proteine, da sie die Doppellipidschicht der Zellmembran durchziehen und somit z. Bsp. als Tunnel für Ionen dienen, sie sind also ein wichtiger Bestandteil der Membran (siehe Abbildung 4). Eine Untergruppe dieser Proteine sind die Transmembranproteine, welche die Doppellipidschicht mit α-helikalen Aminosäuresequenzabschnitten durchziehen. Aufgrund dieser Eigenschaft sind Membranproteine an einem Großteil der biologischen Prozesse beteiligt, dabei handelt es sich unter anderem um die Photosynthese, Nervenerregung oder die Immunantwort der Zellen.

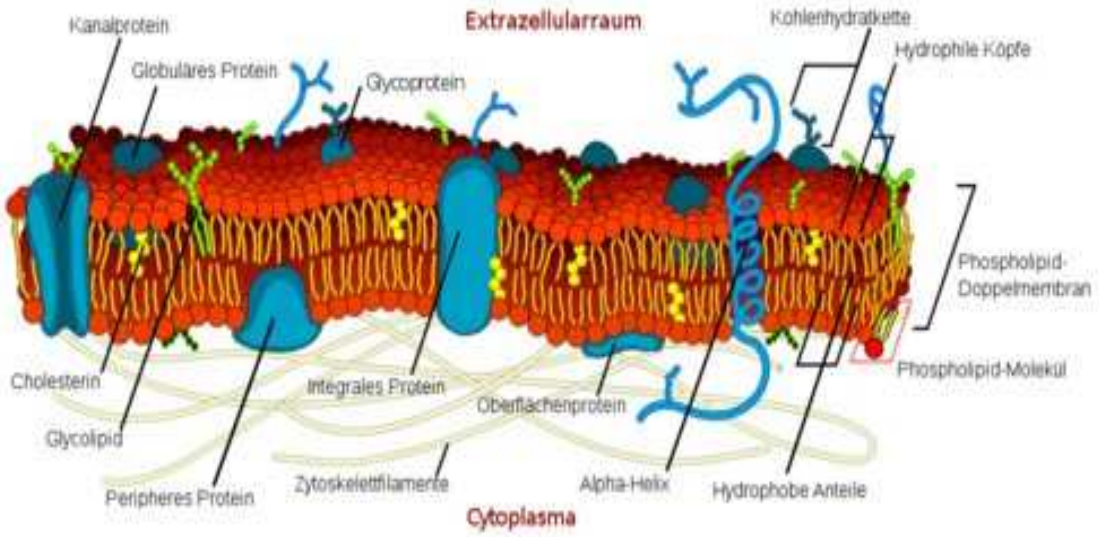

Abbildung 4 Membranproteine
Darstellung verschiedener Membranproteine, eingelagert in der Phospholipid – Doppelmembran.

Da sie die Zellmembran durchziehen, können sie in einzelne Abschnitte unterteilt werden, zum einen sind dies die Transmembranbereiche, welche sich innerhalb der Membran befinden und meist aus Helices bestehen, alles außerhalb wären die nichttransmembranen Bereiche. Zusätzlich gibt es noch die so bezeichneten Transitionbereiche, welche die Übergänge zwischen transmembranen und nichttransmembranen Bereichen darstellen.

2.2 Sequenzmuster

Muster in Proteinen sind Abfolgen von Aminosäuren, die in einer Sequenz lokal konserviert vorhanden sind. Sie besitzen eine wohldefinierte Länge, sprich, die Anfangs- und Endposition sind bekannt und auch, wie viele variable Positionen dazwischen vorhanden sind. Somit haben die in dieser Arbeit verwendeten Muster die Form XZn, X steht für die Anfangsposition, Z für die Endposition und n-1 für die Anzahl der variablen Positionen. Ein Beispiel wäre das GG4 – Muster, ein Glycin steht an der ersten und letzten Position des Musters und dazwischen gibt es drei variable Positionen, auf denen alle 20 kanonischen Aminosäuren auftreten können. Die Quelle der 50 verwendeten Muster (Anhang 1) ist die Arbeit ,,Genomic analysis of membrane protein families: abundance and conserved motifs" von Mark Gerstein et al. [11]. Das gemeinsame Auftreten von zwei (oder mehr) Mustern innerhalb einer Sequenz wird als Co - CO - (*) – Occurence bezeichnet.

2.3 Reguläre Ausdrücke

Reguläre Ausdrücke sind in der Informatik verwendete Zeichenketten, die der Beschreibung von Mengen von Zeichenketten dienen. Somit kann die Suche nach bestimmten Zeichenketten schneller und effektiver gestaltet werden bzw. auch die Erzeugung von Zeichenketten. In dieser Arbeit werden die Ausdrücke verwendet, um bestimmte Muster, welche durch reguläre Ausdrücke beschrieben sind, in Proteinsequenzen zu finden, was als Pattern Matching bezeichnet wird. Diese Ausdrücke sind nach einer festgelegten Syntax aufgebaut, die sich jedoch von Programmiersprache zu Programmiersprache unterscheiden kann. Anhand eines Beispiels soll diese Syntax genauer erläutert werden: Die Muster GCA und GHA sollen als regulärer Ausdruck zusammengefasst und verwendet werden, damit alle vorhandenen Ausprägungen dieser Muster innerhalb bestimmter Proteinsequenzen

gefunden werden. In der Programmiersprache Perl, welche in dieser Arbeit verwendet wurde, wäre der passende reguläre Ausdruck G[CH]A oder G[CH]{1}A und sollten sich in den Sequenzen Abschnitte wie GHA oder GCA befinden, so werden diese mit dem Ausdruck gefunden, Aminosäuren werden innerhalb des regulären Ausdrucks im Ein-Letter-Code dargestellt.

2.4 Ähnlichkeitsbeziehungen

Diese Beziehungen bezeichnen das Zusammenwirken mehrerer Muster in einem Protein. Die einfachste Art einer Ähnlichkeitsbeziehung wäre das gemeinsame Auftreten, sei es nun so, dass sie häufig zusammen auftreten oder auch immer, sprich, wenn Muster X auftritt, dann tritt auch immer Muster Z auf und wenn diese auftreten, impliziert dies möglicherweise das Auftreten von Muster Y. Weitergehende Ähnlichkeitsbeziehungen beziehen sich auf den Einfluss dieser Paarungen auf andere Muster, aber auch auf den Einfluss auf die Struktur und Eigenschaft des gesamten Proteins. Sollten sich solche Ähnlichkeitsbeziehungen erkennen lassen, kann zum einen besser verstanden werden, wie einzelne, wohldefinierte Proteinsequenzfragmente sich auf die Struktur eines Proteins auswirken und zum anderen könnte mit Hilfe mehrerer bekannter Fragmente im Zusammenhang gesehen, versucht werden, in Zukunft ein besseres Drug – Targeting zu gestalten, wenn man sich im Klaren darüber ist, welche Sequenzfragmente gehäuft auftreten und somit bestimmte Funktionen von Proteinen mitbestimmen bzw. prägend für die Struktur sind.

2.5 Pfam - Familien

Für die Untersuchungen in dieser Arbeit wurden 32 Proteinfamilien der Pfam Datenbank ausgewählt. Die Kriterien waren, dass es sich um Transmembranproteine handelt, deren Domänen noch keine Funktion zugeordnet werden konnte, weshalb auf eine genauere Beschreibung verzichtet werden muss [2].

Somit wurden folgende Familien betrachtet:

PF09767, PF09834, PF09842, PF09843, PF09852, PF09858, PF09874, PF09877, PF09878, PF09879, PF09880, PF09881, PF09882, PF09900, PF09913, PF09925, PF09945, PF09946, PF09971, PF09972, PF09973, PF09980, PF09990, PF09991, PF09997, PF10002, PF10011, PF10067, PF10080, PF10081, PF10097 und PF10101.

2.6 Verwendete Datenbanken

2.6.1 Pfam Datenbank 26.0

Die Pfam Datenbank umfasst eine Sammlung von aktuell 13672 Proteinfamilien, welche durch multiple Sequenzalignments und die dazugehörigen Hidden Markov Modelle repräsentiert werden, erzeugt wurde sie vom Wellcome Trust Sanger Institute in Cambridge (UK). Pfam basiert auf der Sequenzdatenbank Pfamseq, welche wiederum auf dem UniProt Release 2011_06 basiert.

Da Proteine im Allgemeinen aus zwei oder mehr funktionellen Regionen, den sogenannten Domänen besteht, können diese Einblicke in die Funktion der Proteine gewähren, was sich die Pfam Datenbank zu Nutze macht. In ihr sind Informationen zur Architektur und Struktur der einzelnen Domänen hinterlegt und auch Informationen darüber, in welchen Organismen sie vorkommen und welche Interaktionen die Domänen ausüben [3]. Genauer gesagt, gibt es auf jeder Familienseite neun Registerkarten:

- Summary – Zusammenfassung,
- Domain organisation – Liste von Domänenorganisation oder -architekturen in denen die Domäne gefunden wurde,
- Alignments – Anzeige oder Download des Alignments,
- HMM logo – Anzeige des HMM Logo,
- Trees – Darstellung des phylogenetischen Baumes der Familie,
- Curation & models – detaillierte Informationen über die PFAM-Familie,
- Species – Vorkommen der Domäne in verschiedenen Spezies,
- Interactions – Verbindungen mit anderen Familien
- Structures – 3D-Strukturen. [1]

In Abbildung 5 ist die Startseite der aktuellen Pfam Release 26.0 dargestellt.

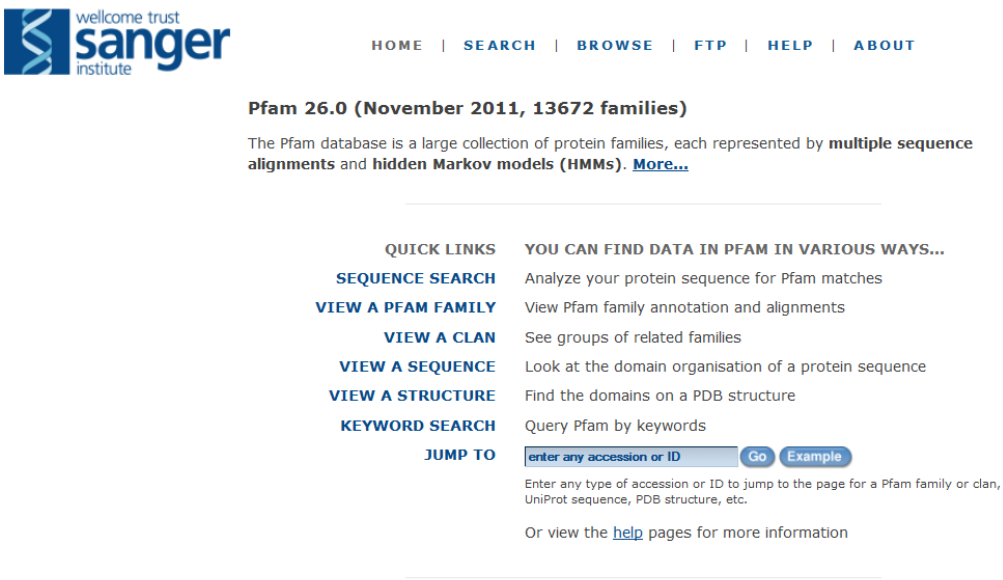

Abbildung 5 Pfam Datenbank

Startseite der Pfam Datenbank mit mehreren auszuwählenden Suchoptionen, es handelt sich um eine sekundäre Datenbank.

Die Proteinfamilien werden in zwei Kategorien eingeteilt, zum einen Pfam – A und zum anderen Pfam – B. Die Einträge in Pfam – A basieren auf manuell erstellten multiplen Alignments repräsentativer Vertreter einer Familie und weisen somit eine hohe Qualität auf. Da Pfam – A nun eine globalisierte Erfassung bekannter Proteine darstellt, von denen noch nicht alle einer charakterisierten Proteinfamilie zugeordnet werden konnten, wurde ein Anhang erstellt, sprich Pfam – B. Hier werden uncharakterisierte Proteinfamilien gespeichert und mit der Zeit möglichst in Pfam – A eingegliedert, sollte sich die Qualität der Einträge erhöhen. [3]

2.6.2 MeMotif

Membranproteine sind wichtig für zahlreiche Prozesse in der Zelle und die zunehmende Anzahl zur Verfügung stehender hochaufgelöster Strukturen macht es erstmals möglich, lokal konservierte strukturelle und funktionelle Motive in α-helikalen Transmembranproteinen.

MeMotif ist ein Projekt der Bioinformatics Group des Biotechnologie Zentrums der TU Dresden und stellt eine Datenbank und Wiki über mehr als 2000 bekannter und neuer rechnerisch vorhergesagter linearer Muster dar, welche in alpha-helikalen

Transmembranproteinen vorkommen und über deren Funktion und Aufbau belegte Aussagen getroffen werden können. Diese sequentiellen Motive wurden von hierarchischen Clustern bekannter α-helikalen Transmembranproteinstrukturen abgeleitet. Dabei werden die Fragmente der Strukturen nach gemeinsamen Mustern von Wasserstoffbrückenbindungen oder ähnlichen Torsionswinkeln der Backbone gruppiert und anschließend jeder Cluster zu einem regulären Ausdruck und einer funktionellen Beschreibung zusammengefasst. In Abbildung 6 ist das Browserfeld von MeMotif dargestellt. Muster der MeMotif Datenbank können für verschiedenste biologische Anwendungen verwendet werden, von der Identifizierung biochemisch gesehen wichtiger funktioneller Reste oder Kandidaten für Mutagenese-Experimente zur Verbesserung der Transmembranprotein – Modellierung [4].

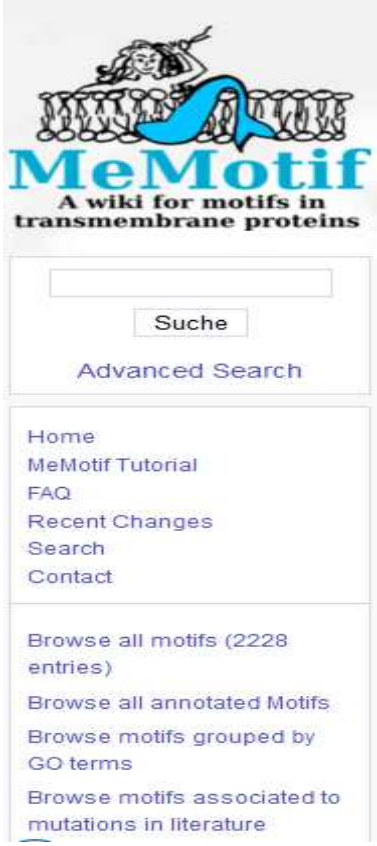

Abbildung 6 MeMotif Datenbank

Bedienfeld der MeMotif Datenbank, in dem man nach bestimmten Mustern suchen kann, sich aber auch eine Liste aller Motive anzeigen lassen kann.

Die Datenbank besteht aus 2228 Motiven, von denen 53 annotiert sind, deren Einträge also detaillierte Informationen zur Funktion des Musters aufweisen und andere Statistiken und Querverweise zu anderen Datenbanken wie Uniprot, Prosite und auch Pfam. Diese Einträge können jederzeit von registrierten Experten editiert oder korrigiert

werden, um den Informationsgehalt zu verbessern und möglichst viele Motive zu annotieren.

2.6.3 ELM

Lineare Muster sind kurze, evolutionär konservierte Bereiche der Sequenz regulatorischer Proteine, welche eine zentrale Rolle in Bezug auf die regulatorischen Funktionen der Zelle einnehmen. Doch obwohl sie so eine große Bedeutung besitzen, ist bisher überraschend wenig über sie bekannt, da sich ihre Entdeckung, egal ob rechnerisch oder experimentell, sehr schwierig gestaltet.

Die Eukaryotic Linear Motif (ELM) Datenbank beinhaltet eine Sammlung experimentell validierter funktioneller Motive in eukaryotischen Proteinen und auch auf ihre Eigenschaft als funktionelle Stelle hin zu untersuchende Kandidaten. Zudem wird ein exploratives Toll zur Verfügung gestellt, mit dessen Hilfe Benutzer nach vermeintlichen linearen Mustern in den von ihnen übermittelten Proteinseuqenzen suchen können. Die aktuelle Version umfasst 1800 kommentierte Muster, welche 170 verschiedene funktionelle Klassen repräsentieren, darunter etwa 500 neue Fälle und 24 neue Klassen. Daneben wurden einige ältere Einträge überarbeitet und Annotationen verbessert und allgemein die Zugänglichkeit der ELM Daten durch eine verbesserte Benutzeroberfläche erleichtert.

Die Einträge werden vom ELM-Konsortium erstellt, welches insgesamt fünf europäische akademische Einrichtungen (EMBL Heidelberg, Universität Rom / Tor Vergata, Universität Bergen, University of Dundee, und die BioInfoBank Institut) und ein Biotechnologie-Unternehmen (Cellzome GmbH) umfasst [5]. In den Einträgen enthalten sind ein regulärer Ausdruck des Motivs, eine umfassende Beschreibung des Motivs, welche Auftreten und Funktion erklärt. Zudem ist hinterlegt, in welchen Organismen, an welcher Stelle der Sequenz, das Motiv gefunden wurde. In Abbildung 7 ist eine Auflistung der ELMs abgebildet.

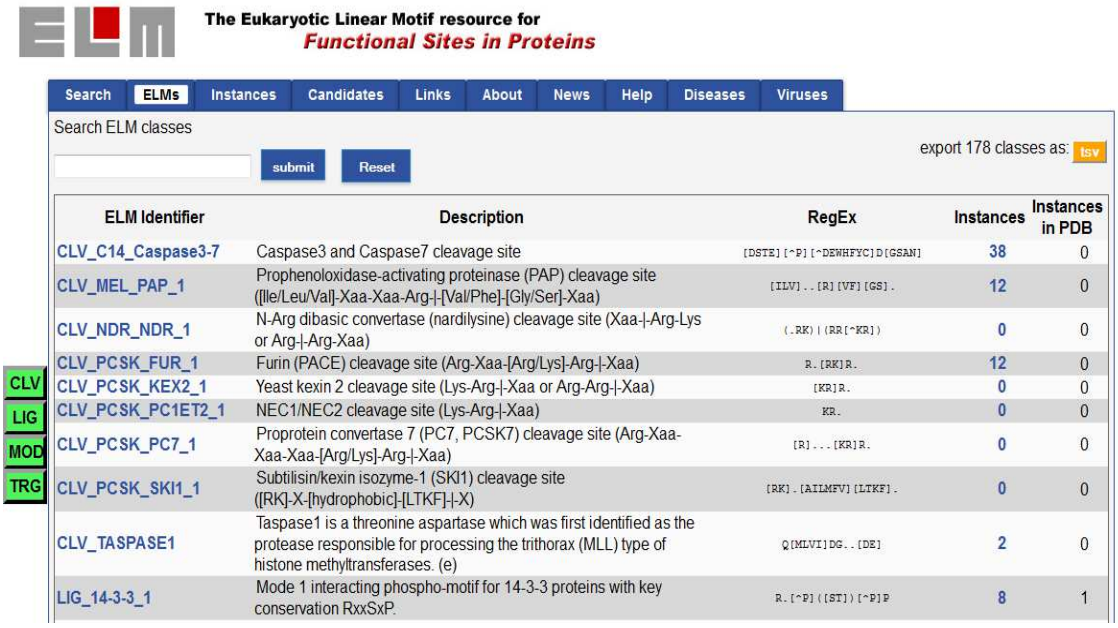

Abbildung 7 Übersicht der ELMs

In dieser Abbildung ist die Seite der ELMs der ELM Datenbank dargestellt, zu erkennen sind die Bezeichnung des Motivs, eine kurze Beschreibung und auch der dazugehörige reguläre Ausdruck.

2.6.4 Prosite

Die Prosite ist eine sekundäre biologische Datenbank, welche 1988 erzeugt wurde.

Die Datenbank enthält Proteinfamilien und Domänen und durch die Analyse dieser konnte festgestellt werden, dass bestimmte Bereiche in der Proteinsequenz besser konserviert als andere sind und daher wichtig für Funktion und Struktur erscheinen. Durch eine gezielte Analyse dieser Bereiche innerhalb einer Proteinfamilie kann eine sequentielle Signatur abgeleitet werden, mit der alle Mitglieder dieser Familie identifiziert werden können. Können Proteinfamilien aufgrund von Unterschieden der Domänen nicht eindeutig durch Signaturen charakterisiert weren, so wird ein Profil für diese in der Datenbank angelegt [6].

Prosite dient somit der Bestimmung von Funktionen uncharakterisierter Proteine und beinhaltet biologisch signifikante Profile und Muster, welche dazu verwendet werden sollen, mit Hilfe geeigneter Algorithmen schnell und zuverlässig analysieren zu können, zu welcher Proteinfamilie eine neue Sequenz gehört, da diese Muster spezifisch für einzelne Proteinfamilien und Domänen sind, meist sind diese in aktiven Zentren oder

Bindungsstellen für Substrate gefunden worden. Mit ScanProsite und ProfileScan stellt Prosite zwei Komponenten zur Verfügung, mit deren Hilfe man übermittelte Proteinsequenzen gegen die Prosite-Datenbank laufen lassen oder mit vorhandenen Proteineinträgen vergleichen lassen kann.

Die Einträge der einzelnen Muster beinhalten eine Beschreibung zu dem Muster mit den dazugehörigen Referenzen von PubMed, Proteine, in denen diese vorkommen und den Konsensus Pattern, also den regulären Ausdruck, der in allen Vertretern übereinstimmend gefunden wurde, als Beispiel soll dafür der Eintrag PDOC00284, dargestellt in Abbildung 8, dienen.

Ursprünglich beinhaltete Prosite nur qualitative Muster, seit 1994 auch quantitative Muster Somit sind mittlerweile 1308 Muster und 1036 Profile integriert [6].


Home ScanProsite ProRule Documents Downloads Links Funding

PROSITE documentation PDOC00284

11-S plant seed storage proteins signature

Description:

Plant seed storage proteins, whose principal function appears to be the major nitrogen source for the developing plant, can be classified, on the basis of their structure, into different families. 11-S are non-glycosylated proteins which form hexameric structures [1,2]. Each of the subunits in the hexamer is itself composed of an acidic and a basic chain derived from a single precursor and linked by a disulfide bond. This structure is shown in the following representation.

```
                +-------------------------+
                |                         |
      xxxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxNGxCxxxxxxxxxxxxxxxxxxxxxxx
                         ********
      <------Acidic-subunit-------------><------Basic-subunit------>
      <-----------------About-450-to-500-residues----------------->
```

'C': conserved cysteine involved in a disulfide bond.
'*': position of the pattern.

Proteins that belong to the 11-S family are: pea and broad bean legumins, rape cruciferin, rice glutelins, cotton β-globulins, soybean glycinins, pumpkin 11-S globulin, oat globulin, sunflower helianthinin G3, etc.

As a signature pattern for this family of proteins we used the region that includes the conserved cleavage site between the acidic and basic subunits (Asn-Gly) and a proximal cysteine residue which is involved in the interchain disulfide bond.

Last update:

June 1994 / Pattern and text revised.

Technical section:

PROSITE method (with tools and information) covered by this documentation:

11S_SEED_STORAGE; PS00305; **11-S plant seed storage proteins signature** (PATTERN)

Consensus pattern:	N-G-x-[DE](2)-x-[LIVMF]-C-[ST]-x(11,12)-[PAG]-D *C is involved in a disulfide bond*
Sequences known to belong to this class detected by the pattern:	ALL
Other sequence(s) detected in Swiss-Prot:	NONE

Abbildung 8 Prosite Eintrag

Hierbei handelt es sich um den Dokumentationseintrag PDOC00284, eine Proteinsignatur. Zu erkennen sind die Beschreibung, der Fundort und der Consensus Pattern, also der reguläre Ausdruck, welcher spezifisch für alle Vertreter der Proteinfamilie ist.

2.7 CD – Hit

CD – HIT steht für Cluster Database at High Identity with Tolerance, wurde von Weizhong Li erstellt und ist mittlerweile ein OpenSource Projekt. Im Folgenden soll die Funktionsweise beschrieben werden: Das Programm CD – Hit benötigt eine Sequenz im FASTA – Format als Input (siehe Abbildung 9) und erzeugt daraus einen Satz von nicht redundanten repräsentativen Sequenzen als Output und eine Clusterdatei, welche die Sequenzgruppierung für jede repräsentative Sequenz dokumentiert. Somit soll die Redundanz verringert werden, aber keine Informationen verloren gehen.

Hierfür wird ein ,,längste Sequenz zuerst“ Listenentfernungs – Algorithmus verwendet, um Sequenzen durch eine bestimmte Schwelle der Identität zu entfernen.

Zusätzlich implementiert der Algorithmus eine sehr schnelle heuristische Methode, um Segmente hoher Identität zwischen Sequenzen zu finden, was viele vollständige Alignments überflüssig macht[7].

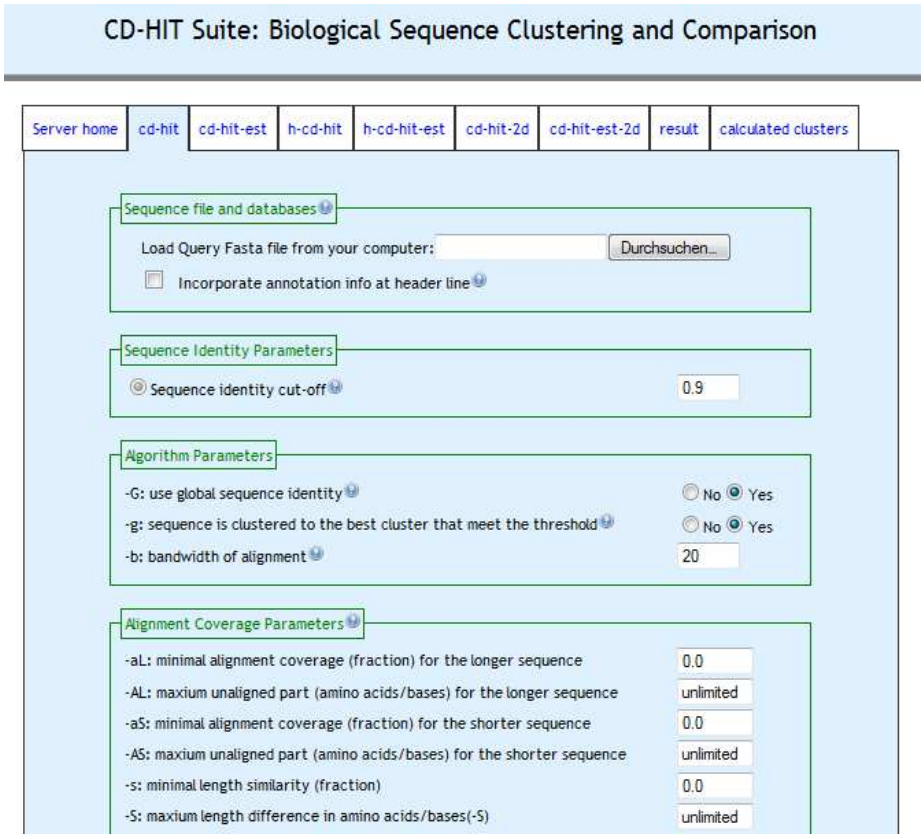

Abbildung 9 Webserver von CD – HIT

Abgebildet ist Startseite des Webservers des Programm CD – Hit, welches vom Benutzer hochgeladene Fasta – Files clustert.

2.8 BlastClust

BlastClust wird dazu verwendet, Protein- oder Nukleotidsequenzen zu clustern. Hierfür beginnt das Programm mit einem paarweisen Alignment unter Verwendung des BLAST Algorithmus im Fall von Proteinen und im Falle von DNA unter Verwendung des Mega BLAST Algorithmus. BlastCLust clustert diese, imdem es eine Sequenz einem Cluster zuordnet, sollte sie mit mindestens einer Sequenz aus diesem Cluster übereinstimmen.

Als Parameter für Proteinsequenzen dienen die Standardeinstellungen von BLAST: Matric BLOSUM62, Gap Opening zählt 11, Gap Extension zählt 1 [8].

2.9 TMHMM 2.0

Der TMHMM 2.0 Server dient der Vorhersage von Transmembranhelices unter der Verwendung von Hidden Markov Modellen, um somit die Funktionen von Transmembranproteinen verhersagen zu können. Das Transmembrane Hidden Markov Model (TMHMM) ist auf dem ,,Center for biological sequence analysis“, kurz CBS, der technischen Universität in Dänemark verankert [1].

Hidden Markov Modelle resultieren aus multiplen Alignments und geben die Wahrscheinlichkeiten für die Übergangszustände innerhalb einer Proteinsequenz an, es handelt sich also um ein stochastisches Modell, in dem ein System, in diesem Fall die Sequenz, durch eine Markov-Kette mit unbeobachteten Zuständen modelliert wird. Die Übergangswahrscheinlichkeiten hängen vom aktuellem Zustand ab und bleiben über die Zeit konstant.

Der TMHMM 2.0 Server benötigt eine Membranproteinsequenz im FASTA-Format als Eingabe, um transmembrane Bereiche detektieren zu können. Als voneinander zu unterscheidene Sequenzbereiche dienen Transmembranbereiche und Bereiche zwischen den transmembranen Helices, welche mit ,,inside“ oder ,,outside“ bezeichnet werden, was sich danach richtet, ob der Sequenzabschnitt sich auf der intrazellulären oder der extrazelluären Seite der Membran befindet [9]. Das TMHMM Ergebnis beinhaltet eine Angabe darüber, auf welchen Positionen welcher Übergangszustand zu finden ist, dies geschieht tabellarisch und grafisch, dargestellt ist dies in Abbildung 10.

TMHMM result

HELP with output formats

```
# Q7Y0D5_ORYSJ Length: 303
# Q7Y0D5_ORYSJ Number of predicted TMHs:  6
# Q7Y0D5_ORYSJ Exp number of AAs in TMHs: 153.89771
# Q7Y0D5_ORYSJ Exp number, first 60 AAs:  22.75312
# Q7Y0D5_ORYSJ Total prob of N-in:        0.99861
# Q7Y0D5_ORYSJ POSSIBLE N-term signal sequence
Q7Y0D5_ORYSJ    TMHMM2.0        inside      1     6
Q7Y0D5_ORYSJ    TMHMM2.0        TMhelix     7    29
Q7Y0D5_ORYSJ    TMHMM2.0        outside    30    76
Q7Y0D5_ORYSJ    TMHMM2.0        TMhelix    77    99
Q7Y0D5_ORYSJ    TMHMM2.0        inside    100   110
Q7Y0D5_ORYSJ    TMHMM2.0        TMhelix   111   133
Q7Y0D5_ORYSJ    TMHMM2.0        outside   134   147
Q7Y0D5_ORYSJ    TMHMM2.0        TMhelix   148   165
Q7Y0D5_ORYSJ    TMHMM2.0        inside    166   244
Q7Y0D5_ORYSJ    TMHMM2.0        TMhelix   245   267
Q7Y0D5_ORYSJ    TMHMM2.0        outside   268   276
Q7Y0D5_ORYSJ    TMHMM2.0        TMhelix   277   299
Q7Y0D5_ORYSJ    TMHMM2.0        inside    300   303
```

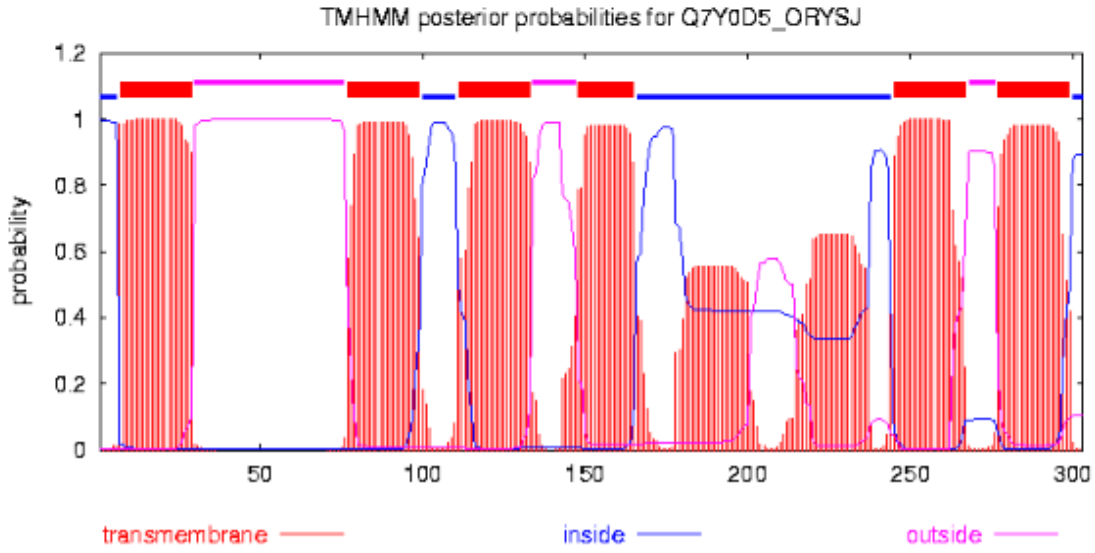

Abbildung 10 Ergebnis TMHMM. Ausgabe des TMHMM 2.0 Servers in tabellarischer und grafischer Form.

2.10 Weblogo

Ein Weblogo ist ein durch eine web-basierte Anwendung erstelltes Sequenzlogo von Nuklein- oder Aminosäuren. Es stellt also ein Visualisierungstool für ein multiples Alignment dar, durch welches der Informationsgehalt des Alignments angehoben wird.

Dabei werden mehrere Sequenzen der gleichen Länge auf ihre einzelnen Positionen hin untersucht. In jedem Logo werden die Ausprägungen der Sequenzen innerhalb von Säulen für jede Position spezifisch dargestellt. Eine große Säule steht für eine stark konservierte Position, da die Höhe der einzelnen Ausprägung innerhalb der Säule für eine große Wahrscheinlichkeit des Auftretens dieser Amino- bzw. Nukleinsäure an der jeweiligen Position steht. Die Farbe der auftretenden Aminosäuren richtet sich nach ihren chemikalischen Eigenschaften, polare werden grün, basische blau, saure rot und hydrophobe Aminosäuren werden schwarz dargestellt. In Abbildung 11 ist ein WebLogo für CAP Binding Sites als Beispiel dargestellt.

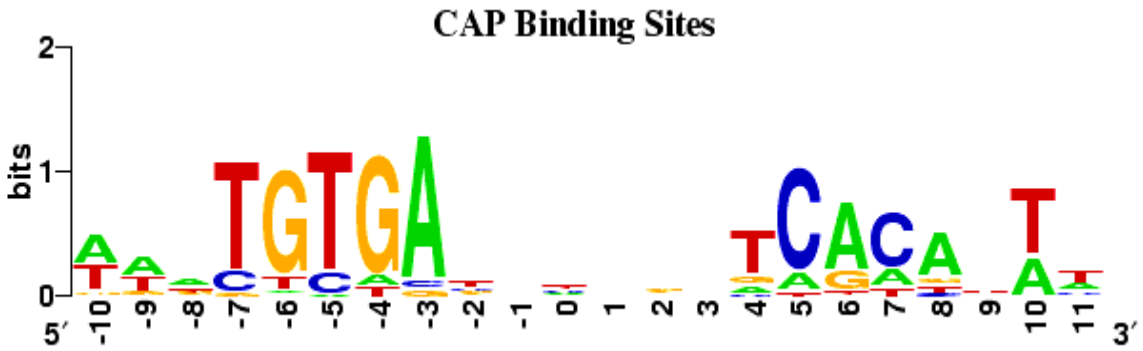

Abbildung 11 Beispiel für ein WebLogo

Darstellung eines WebLogos für CAP Binding Sites, in dieser Form wurden WebLogos für zu untersuchende Musterpaarungen erzeugt.

WebLogos beinhalten somit folgende Informationen in nur einer Grafik:

- die Konsensussequenz der übermittelte Sequenzen,
- die Abfolge der Reste und deren relative Häufigkeit an jeder Position und somit den Grad der Konservierung des jeweiligen Restes
- einen Anfangs- und Endpunkt und mögliche andere signifikante Positionen [1].

2.11 Verwendete Programme

In der vorliegenden Arbeit wurden selbst programmierte Programme verwendet, aber auch Programme anderer Entwickler, was im Folgenden auch erwähnt wird.

MusterPrax12

Mit Hilfe dieses Programms wurden die Proteinsequenzen in TM- und nTM – Bereiche gesplittet und darin die Muster gesucht. Dafür benötigte das Programm den Pfad der FASTA – Sequenzen und den Pfad des TMHMM – Files, durch Zuordnung der Ergebnisse des TMHMM zu den Sequenzen konnten diese gesplittet werden und die darin enthaltenen Muster wurden in eine TXT – Datei geschrieben und gezählt.

grustentier.bio.java.small

Das Programm mit dem Namen grustentier.bio.java.small wurde von Master of Science Steffen Grunert entwickelt und diente der Untersuchung der Transitionbereiche auf Auftreten der einzelnen Motive. Hierfür schaute das Programm in dem TMHMM – File für die als Input angegebenen FASTA – Sequenzen nach, welche Muster auf welchen Positionen vorkommen. Befand sich die Startposition des Musters im transmembranen und die Endposition im nichttransmembranen Bereich oder umgekehrt, so wurde das Muster als im Transitionbereich vorkommend ausgegeben und stand für weitere Untersuchungen zur Verfügung.

CoCoOccurence

Mit denen durch die Programme MusterPrax12 und grustentier.bio.java.smallerstellten Daten konnte das Programm CoCoOccurence nun untersuchen, ob zwei Muster zusammen in einer Proteinsequenz vorkommen oder nicht. Hierfür benötigt das Programm die transmembranen Abschnitte, dann die nichttransmembranen Abschnitte und für die Transitionbereiche wurden die Muster herangezogen, die in dem Bereich bereits einzeln gefunden wurden. Durch Erstellen von 50x50 Matrizen für jede Proteinsequenz und einmal für alle 32 Proteinfamilien globulär gesehen, jeweils separat für die verschiedenen Bereiche, konnte das Programm hochzählen, wie oft bestimmte Muster miteinander vorkommen.

Perl Modul Regexp-Genex-0.07 von Brad Bowman

Dieses Perl Modul [10] ermöglichte es, aus einem regulären Ausdruck Strings zu erzeugen, auf die eben jener Ausdruck matchen würde. Die Strings konnten in einer Datei gespeichert werden und standen so für weitere Untersuchungen zur Verfügung.

3 Methoden

3.1 Auftreten der Gersteinmuster in Transmembranproteinen

Wie in den Grundlagen bereits erwähnt, dienten als Basis der Untersuchungen 32 Transmembranproteinfamilien der Pfam Datenbank, von deren Domänen noch keine Funktion bekannt ist. Alle Proteinsequenzen im Fasta - Format wurden mittels CD-Hit verglichen und geclustert bei einer Sequenzidentität von 60%. Daraufhin folgte ein Clustering auf 25% Sequenzidentität mittels BlastClust, was dazu beitragen soll, dass Redundanzen reduziert werden. Anschließend wurden diese geclusterten Fastasequenzen an den TMHMM 2.0 Server übergeben, um die transmembranen Helices vorhersagen zu können und das Ergebnis wurde als HTML – File gespeichert. Da in der Arbeit die transmembranen, nichttransmembranen Bereiche und auch die Transitionbereiche getrennt voneinander untersucht werden sollten, mussten die Proteinsequenzen der 32 Familien zunächst mit Hilfe der jeweiligen TMHMM - Resultate aufgesplittet werden, wozu das Programm MusterPrax12 diente. Für die Analyse der Transitionbereiche diente das Programm grustentier.bio.java.small.

Als erstes wurde untersucht, welche der 20 kanonischen Aminosäuren wie oft in den einzelnen Bereichen der Transmembran und im nichttransmembranen Bereich vorkommen, verwendet wurde hierfür das Programm MusterPrax12.

Die ersten Untersuchungen in den einzelnen Bereichen, welche sich mit den Mustern befassten, analysierten das Auftreten der einzelnen Muster in diesen. Dieses Auftreten wurde mit Hilfe des Programms MusterPrax12 gezählt und mit den daraus resultierenden Daten wurden relative Häufigkeiten der Muster anhand folgender Formel errechnet:

Formel 1: Relative Häufigkeit

$$relative\ Häufigkeit = \frac{absolute\ Häufigkeit\ des\ Musters\ X\ \ für\ Bereich\ Y}{Anzahl\ aller\ Muster\ \ für\ Bereich\ Y}$$

Somit kann analysiert werden, welche Muster auch allein häufig in Transmembranproteinen vorkommen und möglicherweise eine wichtige Aufgabe betreffs Struktur oder Funktion des Proteins erfüllen.

3.2 Untersuchung der Co – Co – Occurence

Um mögliche Ähnlichkeitsbeziehungen zwischen den Mustern erkennen zu können, wurde im nächsten Schritt untersucht, welches Muster zusammen mit sich selbst oder einem anderen Muster innerhalb der selben Sequenz auftritt. Durch die Analyse der dafür durch das Programm CoCoOccurence erstellten 50x50 - Matrizen für die Co – Co – Occurence konnte analysiert werden, welche Paarungen besonders häufig auftreten, dies geschah zum einen global gesehen für alle 32 Familien zusammengefasst und zum anderen für jede einzelne Proteinfamilie isoliert betrachtet. Für den Vergleich der selben Paarungen in unterschiedlichen Bereichen wurden zudem WebLogos erstellt, um somit möglicherweise klären zu können, in wie weit einzelne, auf den Zwischenpositionen konservierte Aminosäuren, Einfluss auf die Funktion oder das bloße Vorhandensein dieser Paarungen nehmen. Um nicht nur die bloße Anzahl des Auftretens der Paarungen als Grundlage zu haben, wurde versucht, eine Normierung durchzuführen. Hierfür wurde für jede Paarung geschaut, wie oft sie aufgrund des einzelnen Auftretens der Muster in den Proteinsequenzen theoretisch auftreten kann und wie oft sie wirklich auftrat, es handelt sich somit also um eine relative Häufigkeit, mit der versucht wurde, Paarungen zu identifizieren, bei denen, wenn ein Muster auftritt, fast zwangsläufig auch das zweite auftritt. Die dafür verwendete Formel lautet:

Formel 2: Relative Häufigkeit für das Auftreten als Paarung

$$relative\ Häufigkeit = \frac{Anzahl\ der\ Paarung\ XZ\ \ für\ Bereich\ Y}{Anzahl\ Muster\ X\ \ für\ Bereich\ Y\ x\ Anzahl\ Muster\ Z\ \ für\ Bereich\ Y}$$

3.3 Literatur- und Datenbankrecherche

Da anhand der Co – Co - Occurence nur ausgesagt werden kann, ob bestimmte Muster gehäuft miteinander auftreten oder nicht, jedoch keine Aussage über den Grund dafür getroffen werden kann, musste recherchiert werden, welchen einzelnen Mustern bereits in der Literatur oder auf Datenbanken eine Funktion zugeordnet werden konnte. Hierfür wurden zum einen Muster mit beschriebener Funktion in der Literatur gesucht und zum anderen wurden Muster und Pattern auf den Seiten der Prosite, ELM und MeMotif

durchsucht. Prosite und ELM wurden auf Muster durchsucht, welche in Helices oder Membranen vorkommen, um eine möglichst große Datenbasis zu erlangen, von MeMotif wurden alle annotierten Muster verwendet.

3.4 Erzeugung der Musterstrings

Mit Hilfe des Perl- Moduls Regexp-Genex-0.07 von Brad Bowman wurden alle Ausprägungen der Datenbankenmuster erzeugt und in diesen dann die 50 Muster aus der Arbeit von Gerstein et al. gesucht. Somit kann diesen 50 Mustern möglicherweise eine Funktion zugeordnet werden, sollten sie sich in den Mustern aus den Datenbanken finden lassen.

3.5 Zuordnung von Funktionen

Der letzte Schritt bestand darin, Funktionen der Muster mit der Co – Co - Occurence zusammenzubringen, also zu analysieren, welche Paarung aufgrund der möglichen Funktion der einzelnen Muster eine bestimmte Aufgabe im Protein erfüllt. Hierfür wurden die häufigsten Paarungen im Transmembranbereich untersucht, wobei die einzelnen Partner in unterschiedlichen Helices auftreten mussten, zudem wurden auch die häufigsten Paarungen im Transitionbereich untersucht und ausgewählte Paarungen, die in diesen Bereichen nur in bestimmten Familien gehäuft auftreten.

4 Ergebnisse

4.1 Häufigkeitsanalysen

4.1.1 Auftreten der Aminosäuren

Um beurteilen zu können, welche Aminosäuren besonders häufig in den transmembranen und nichttransmembranen Bereichen auftauchen, wurde das Auftreten dieser gezählt. Dies diente zum einen der Beurteilung der Zwischenpositionen der Muster und zum anderen der Erklärung, weshalb bestimmte Muster verstärkt in bestimmten Bereichen vorkommen. Dargestellt ist diese Zählung in der Tabelle 2.

Tabelle 2 Auftreten der Aminosäuren

Liste der 20 kanonischen Aminosäuren, sortiert nach Anzahl in den durchsuchten Proteinsequenzen, zum einen für den transmembranen Bereich und zum anderen für den nichttransmembranen Bereich.

TM – Bereich		nTM – Bereich	
Aminosäure	Anzahl	Aminosäure	Anzahl
Leucin	106749	Alanin	42138
Alanin	73501	Glycin	39699
Valin	59343	Leucin	39160
Isoleucin	51209	Arginin	31053
Glycin	44169	Serin	30127
Phenylalanin	40096	Valin	26336
Serin	26111	Prolin	25138
Threonin	24794	Asparaginsäure	22765
Tyrosin	18423	Glutaminsäure	22677
Tryptophan	14694	Threonin	22487
Methionin	13781	Isoleucin	18110
Prolin	12602	Lysin	16527
Arginin	7047	Glutamin	15101
Asparagin	6106	Phenylalanin	15077
Cystein	6038	Tyrosin	12820
Glutamin	5413	Asparagin	12562
Histidin	4319	Methionin	9212
Glutaminsäure	3503	Histidin	9005
Lysin	2659	Tryptophan	7833
Asparaginsäure	2562	Cystein	2880

Anhand der Tabelle 2 ist sehr gut zu erkennen, dass Leucin und Alanin die mit Abstand am häufigsten auftretenden Aminosäuren im transmembranen Bereich sind, im nichttransmembranen Bereich sind sie auch unter den am häufigsten auftretenden. Insgesamt gesehen lässt sich aussagen, dass die Aminosäuren unterschiedlich oft in bestimmten Bereichen auftreten, was mit ihren physikochemischen Eigenschaften zu begründen ist. So ist es beispielsweise logisch, dass sich hydrophobe Aminosäuren vor allem in der Membran und hydrophile Aminosäuren vor allem außerhalb der Membran befinden. Dass einige Aminosäuren in beiden Bereichen oft auftreten, hängt damit zusammen, dass es sich um Aminosäuren, die auch in der Natur generell am häufigsten in Proteinsequenzen vorkommen.

4.1.2 Auftreten der einzelnen Muster

Nach dem Zählen des Auftretens der einzelnen Muster in den transmembranen und nichttransmembranen Bereichen (Anhang 1) und der Berechnung der relativen Häufigkeit des Vorkommens dieser Muster in dem jeweiligem Bereich, stellten sich folgende der insgesamt 50 untersuchten und gezählten Muster, sortiert in Tabelle 3, als am jeweils häufigsten auftretend heraus:

Tabelle 3 Häufigste Muster im TM – und nTM -Bereich

Liste der jeweils fünf am häufigsten auftretenden Muster in den beiden Bereichen, in der ersten Spalte steht jeweils das gefundene Muster, in der zweiten die Anzahl der Treffer und in der dritten Spalte ist die relative Häufigkeit angegeben.

Häufigste Muster					
TM – Bereich			nTM – Bereich		
Muster	Anzahl	rel. HF	Muster	Anzahl	rel. HF
VL4	6677	0,0432	AA2	4655	0,0376
AL6	6605	0,0428	AA3	4646	0,0375
GL1	6405	0,0415	LS1	3857	0,0311
AA2	5852	0,0379	AA7	3741	0,03
GL3	5818	0,0377	AL6	3605	0,0292

Schon bei dieser relativ kleinen Auswahl ist zu erkennen, dass bestimmte Muster unterschiedlich oft in den verschiedenen Bereichen vorkommen, was damit zu erklären ist, dass die beinhaltenen Aminosäuren unterschiedliche physikochemische Eigenschaften besitzen, also das Muster insgesamt gesehen entweder einen hydrophoben oder hydrophilen Charakter besitzt, was dazu führt, dass sie in bestimmten Bereichen bevorzugt vorkommen. Jedoch ist am Beispiel des AA2 Musters auch zu erkennen, dass manche Muster immer gehäuft auftreten, dies kann zum einen daran liegen, dass sie eine bestimmte Funktion im Protein innehaben aber auch daran, dass sie in der Natur oft vorkommende Aminosäuren beinhalten, was bei dem in AA2 enthaltenen Alanin der Fall ist, weshalb Muster, welche Alanin oder auch Leucin enthalten, rein statistisch gesehen, gehäuft vorkommen. In der Abbildung 12 ist noch einmal zu sehen, wie die einzelnen Muster in den beiden Bereichen prozentual gesehen auftreten.

Für den Transitionbereich wurden nur Muster mit einer Mindestlänge von drei Aminosäuren gezählt, da kleinere Muster eher weniger aussagekräftig für diesen Bereich sind, der im Regelfall um die 5 Aminosäuren in jedem Bereich umfassen kann, also insgesamt eine Länge von 10 Aminosäuren besitzt. Die häufigsten zehn Muster wurden in der Tabelle 4 zusammengefasst.

Tabelle 4 Muster im Transitionbereich

Dargestellt sind die zehn am häufigsten auftretenden Muster im Transitionbereich.

Muster	Anzahl
AL6	2576
LF10	2554
AA7	2256
LF9	2191
LF8	2027
AG7	1929
GA7	1760
LG5	1740
AG5	1305
GG7	1260

Wie auch im transmembranen und nichttransmembranen Bereich kommt das Muster AL6 besonders häufig vor und im allgemeinen kommen im Transitionbereich vor allem Muster mit einder Mindestlänge von sechs Aminosäuren vor, was der Struktur dienlich erscheint.

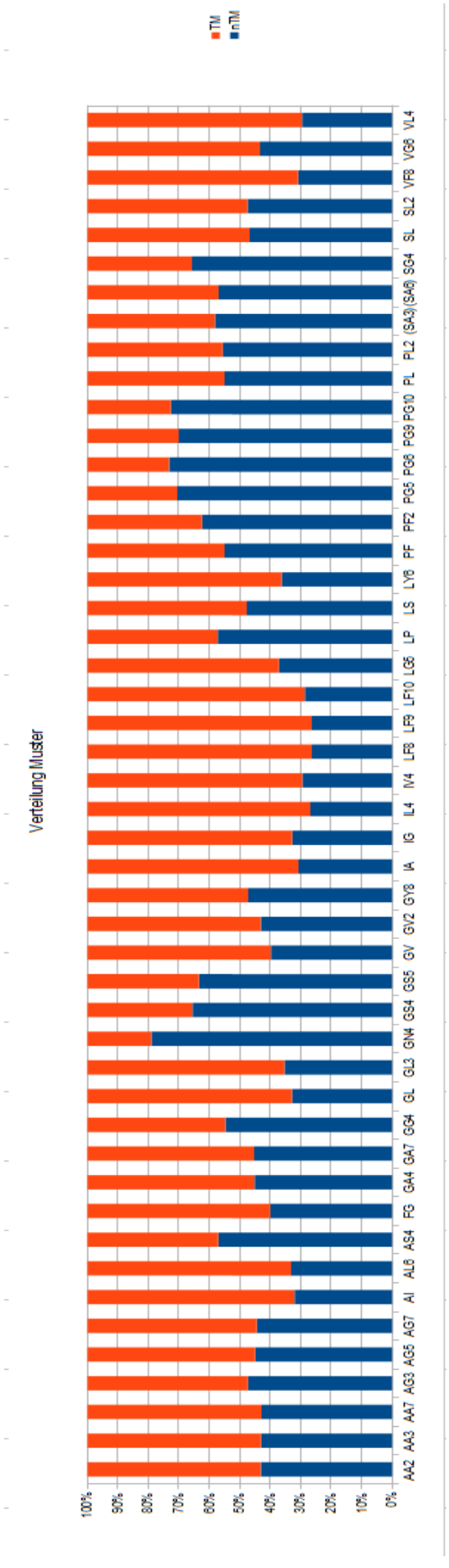
Verteilung Muster
TM
nTM
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
AA2
AA3
AA7
AG3
AG5
AG7
AI
AL6
AS4
FG
GA4
GA7
GG4
GL
GL3
GN4
GS4
GS5
GV
GV2
GY8
IA
IG
IL4
IV4
LF8
LF9
LF10
LG6
LP
LS
LY6
PF
PF2
PG5
PG6
PG9
PG10
PL
PL2
(SA3)(SA6)
SG4
SL
SL2
VF8
VG6
VL4

Abbildung 12 Verteilung der Muster

Zu erkennen ist, in welchem Bereich, prozentual miteinander verglichen, die Muster wie häufig auftreten. Zum Beispiel ist zu sehen, dass GN4 viel öfter im nTM – Bereich (blauer Teil der Säulen) als im TM – Bereich (orange gefärbter Teil der Säulen) auftritt.

Auch durch diese Analyse zeigt sich, wie unterschiedlich häufig die einzelnen Muster in TM – und nTM – Bereichen auftreten bzw. wie repräsentativ sie für die einzelnen Bereiche sind. So ist beispielsweise zu erkennen, dass GN4 viel öfter im nTM – Bereich auftritt als im TM – Bereich, bei LF9 dagegen ist das Gegenteil der Fall. Insgesamt betrachtet, scheint es zudem der Fall zu sein, dass Muster mit einer Länge von mehr als sechs Aminosäuren viel öfter im transmembranen Bereich als im nTM – Bereich auftreten, was zum einen daran liegen kann, dass Helixabschnitte länger als outside und inside Bereiche sind oder dass diese Muster eine tragende Funktion innerhalb der Helices besitzen.

4.2 Co – Co – Occurence

Mit Hilfe des Programms CoCoOccurence wurden 50x50 Matrizen für die 32 Proteinfamilien für die drei Bereiche TM, nTM und Transition erstellt, ein Beispiel ist in Tabelle 5 für die Familie PF09874 im transmembranen Bereich dargestellt. Diese Matrizen stellten die Basis für alle weiteren, die Co – Co – Occurence betreffenden, Untersuchungen in dieser Arbeit dar, da sie aufzeigen, welche Paarungen in welcher Proteinfamilie und für alle 32 Familien global betrachtet, wie oft auftreten. Neben diesen Matrizen wurde zudem jeweils eine Textdatei mit den jeweiligen Ausprägungen erstellt, welche der genaueren Untersuchung dienten und auch für die Erstellung der in dieser Arbeit verwendeten WebLogos verwendet wurden.

Tabelle 5 Co – Co – Occurence Tabelle

Dargestellt ist eine Tabelle der Co - Co – Occurence für eine einzelne Proteinfamilie. Zu erkennen ist, welche Muster miteinander wie oft auftreten, dabei wird für eine Paarung nur einmal hochgezählt, also entweder für die Paarung XZ (Muster X und Z) oder für die Paarung ZX, welche Kombination hochgezählt wird, hängt von der alphabetischen Suche des Programms ab.

Muster	AI	AA2	GG7	IG	LP	AY6	LF8	LF10	PG10	SL	AA3	AS4	IA	PL	VL4	AG4	FG	GA4	IL4	VF8	AG5	AL6	GL2	LF9	SA3	SA6	AA7	LS	PL2	VG6	GL	GL3	GS5	GG7	IG2	IV4	LG5	PG6	SL2	GV	GG4	GV2	GY8
AI	1	2	1	2	4	2	6	8	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AG5	0	1	2	0	9	1	1	5	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GA7	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IG	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LP	0	0	5	2	6	2	0	0	4	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LY6	0	0	1	3	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF8	0	0	2	12	3	6	6	9	0	16	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	2	14	9	4	15	10	2	23	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PG10	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SL	0	0	3	11	0	1	0	0	0	5	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA3	0	2	0	2	2	2	5	6	0	6	1	1	9	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AS4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IA	2	3	2	12	5	6	15	15	0	14	0	1	5	4	15	0	0	1	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PL	0	0	1	4	1	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VL4	3	8	3	11	14	2	7	20	3	17	0	0	0	2	9	0	3	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AG3	1	2	1	0	4	0	0	2	1	2	0	0	0	0	3	0	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FG	1	2	1	1	4	0	2	5	1	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GA4	1	2	1	1	4	0	1	4	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IL4	4	6	8	22	16	9	24	21	2	22	8	1	26	9	18	0	3	2	20	15	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VF8	1	2	3	7	6	2	5	5	1	7	0	0	0	2	8	0	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AG5	0	0	0	0	0	0	2	2	0	3	1	1	1	0	1	0	0	1	1	0	0	2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AL6	5	4	1	7	6	4	16	22	1	13	6	1	12	1	13	1	3	3	14	2	0	3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GL2	6	4	2	13	8	2	11	22	2	17	0	0	7	0	18	2	4	4	12	7	0	12	9	5	3	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
LF9	2	2	1	7	3	4	15	16	0	13	7	2	14	2	9	0	1	1	17	3	0	12	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SA3	1	2	1	1	4	0	3	6	1	6	1	1	2	0	6	1	1	2	3	2	0	3	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SA6	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AA7	2	1	0	1	1	2	4	4	0	2	2	0	4	0	3	0	0	1	3	0	1	5	3	3	0	0	0	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0
LS	7	2	4	22	7	6	25	28	1	27	6	3	22	6	19	1	1	3	38	11	4	15	22	16	5	0	0	15	2	1	0	0	0	0	0	0	0	0	0	0	0	0
PL2	2	5	2	0	9	1	1	5	2	5	2	0	3	0	8	2	2	2	6	2	0	4	4	2	2	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0
VG6	1	3	1	0	5	1	1	3	1	3	2	0	3	0	5	1	1	1	4	1	0	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GL	3	2	1	2	4	1	4	7	1	4	0	0	2	0	7	1	1	3	3	2	1	5	6	2	2	0	2	6	2	1	0	3	2	1	2	6	2	1	4	0	0	0
GL3	2	0	1	11	1	2	9	12	0	9	1	0	8	2	10	0	1	2	13	5	1	7	14	5	1	1	2	17	0	0	0	1	3	1	3	9	5	0	11	0	0	0	1
GS5	1	2	1	2	4	0	0	3	1	4	0	0	1	0	6	1	1	2	3	2	1	2	7	0	1	0	1	5	2	1	0	0	0	1	2	3	0	1	5	0	0	0	0
GG7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
IG2	1	3	1	3	5	1	4	8	1	5	3	0	5	1	8	1	2	2	8	1	1	7	5	4	1	1	2	5	3	2	0	0	0	0	0	12	0	1	7	0	0	0	0
IV4	6	8	2	11	12	6	20	30	2	19	12	1	22	3	24	2	4	4	26	3	2	24	14	18	4	2	7	22	8	6	0	0	0	0	0	12	0	2	26	0	0	0	0
LG5	0	1	2	6	3	3	6	3	0	4	3	0	9	3	6	0	0	2	17	4	2	5	2	5	0	0	3	11	1	1	0	0	2	2	4	9	2	0	7	0	0	0	0
PG6	1	2	1	0	4	0	0	2	1	2	0	0	0	0	3	1	1	1	2	1	0	1	2	0	1	0	0	1	2	1	0	0	0	0	0	0	0	0	2	0	0	0	0
SL2	4	7	3	14	12	5	13	21	2	21	9	1	21	3	25	2	2	3	27	8	1	14	19	13	4	0	4	27	7	5	0	0	0	0	0	0	0	0	8	0	0	0	0
GV	2	5	2	6	9	1	4	12	2	10	3	0	8	1	19	2	2	5	10	5	1	8	11	5	4	0	2	11	5	3	5	6	4	1	5	14	4	2	14	2	3	0	0
GG4	1	2	1	2	4	0	0	3	1	4	0	0	1	0	5	1	1	1	3	2	0	1	6	0	1	0	0	4	2	1	1	2	2	0	1	2	0	1	5	0	0	0	0
GV2	2	6	2	8	10	2	9	19	2	12	6	0	13	2	24	2	3	6	15	5	1	14	10	10	5	1	3	12	6	4	6	7	3	1	8	24	6	2	17	15	2	6	0
GY8	0	0	1	2	1	1	2	0	0	1	0	0	2	1	0	0	0	0	6	2	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0

Nach Analyse dieser Matrizen für alle drei Bereiche kristallisierten sich diejenigen Paarungen heraus, welche am häufigsten auftreten (vergleiche Tabelle 6).

Tabelle 6 Häufigste Musterpaarungen

Liste der jeweils zehn am häufigsten auftretenden Muster in den TM -, nTM – und Transitionbereichen, in der ersten Spalte steht jeweils die gefundene Paarung und in der zweiten Spalte die Anzahl der Treffer.

Häufigste Musterpaarungen in den einzelnen Bereichen					
nTM		TM		Transition	
Paarung	Anzahl	Paarung	Anzahl	Paarung	Anzahl
AA2AA3	24089	AA2AA3	42155	AA2AA3	25969
AA3AA7	20012	AA2AL6	39043	AA2GL1	18855
AA2AA7	19689	AA3AL6	38006	AA3GL1	18449
AA3AL6	16060	AL6GL1	35247	AA3AG4	18096
AA2AL6	15897	VL4AL6	33722	GL1GL3	17967
AA3AG4	15705	AA2GL1	33399	AA2GA4	17934
AA2AG4	15570	AL6GL3	33387	GL1GL2	16393
AA3GA4	15051	AA3GL1	32266	AA2AL6	15911
AA2GA4	15011	VL4GL1	32082	AA3AL6	15806
AA3AG5	14430	AA2GL3	31753	AL6GL1	15794

Vergleicht man das Auftreten der häufigsten zehn Paarungen von theoretisch 1225 möglichen ((50x50)/2, da beispielsweise die Paarung AA2AA3 der Paarung AA3AA2 entspricht) in den einzelnen Bereichen, so erkennt man, dass bestimmte Paarungen wie AA2AA3, AA2AL6 und AA3AL6 in allen Bereichen auftreten und meist am häufigsten. Sieht man sich nun die Ausprägungen mit Hilfe von WebLogos an, so ist zu erkennen, dass es sich wahrscheinlich oft um Paarungen handelt, bei dem das kürzere Muster im größeren gefunden wird, da zunächst nicht beachtet wird, ob die einzelnen Partner möglicherweise an gleicher Position gefunden wurden, da sich keine Aussage darüber treffen lässt, welches Muster denn nun wichtiger als ein anderes ist, also das kleinere, welches im größeren enthalten ist oder eben das größere an sich (siehe Abbildung 13).

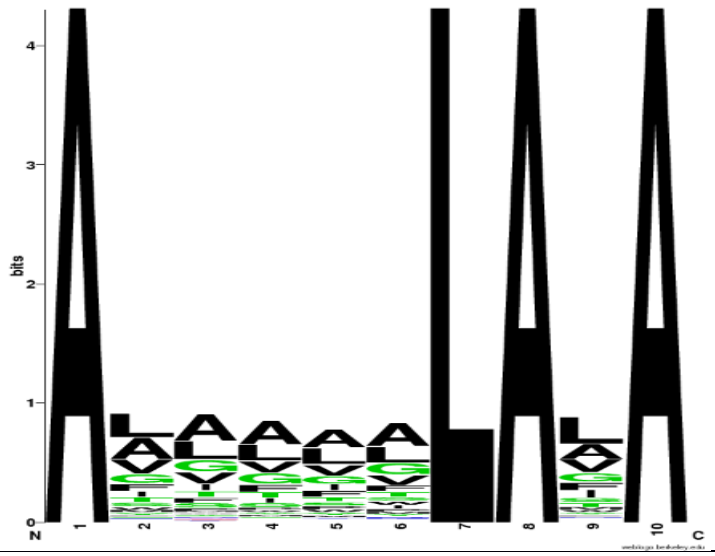

Abbildung 13 WebLogo für die Paarung AL6AA2 im transmembranen Bereich

Das WebLogo für AL6AA2 zeigt, dass auf allen Positionen am häufigsten Alanin oder Leucin auftreten und das AA2 Muster beispielsweise auch einfach auf den ersten drei Positionen des AL6 Musters gefunden werden konnte.

Für genauere Untersuchungen , vor allem im TM – Bereich sollte daher geschaut werden, dass die einzelnen Partner in verschiedenen Helices auftreten, da solch eine Ähnlichkeitsbeziehung aussagekräftiger in Hinblick auf ihren Einfluss auf die Struktur des Proteins ist. Zudem sind Alanin und Leucin die am häufigsten auftretenden Aminosäuren in den Sequenzen, wodurch Muster, welche diese beinhalten oder größtenteils daraus bestehen auch häufig vorzufinden sind, vor allem bei so kurzen Mustern wie AA2 und AA3. Kommen die Partner in unterschiedlichen Helices vor, so lassen sich auch Aussagen darüber treffen, wie diese untereinander Interaktionen verwirklichen.

Jedoch sind auch Paarungen identifizierbar, die vor allem in einem bestimmten Bereich

auftreten, wie AA3AA7 im nichttransmembranen Bereich, GL1GL3 im Transitionbereich und VL4AL6 im transmembranen Bereich. Der Grund hierfür ist in den verschiedenen Ausprägungen der Muster, also den Zwischenpositionen zu suchen, welche die Eigenschaften der Muster bestimmen. Betrachtet man WebLogos für die gleichen Paarungen in unterschiedlichen Bereichen, so lässt sich feststellen, dass sie andere Zwischenpositionen beinhalten und womöglich nur deshalb in den Bereichen an ihren Positionen auftreten.

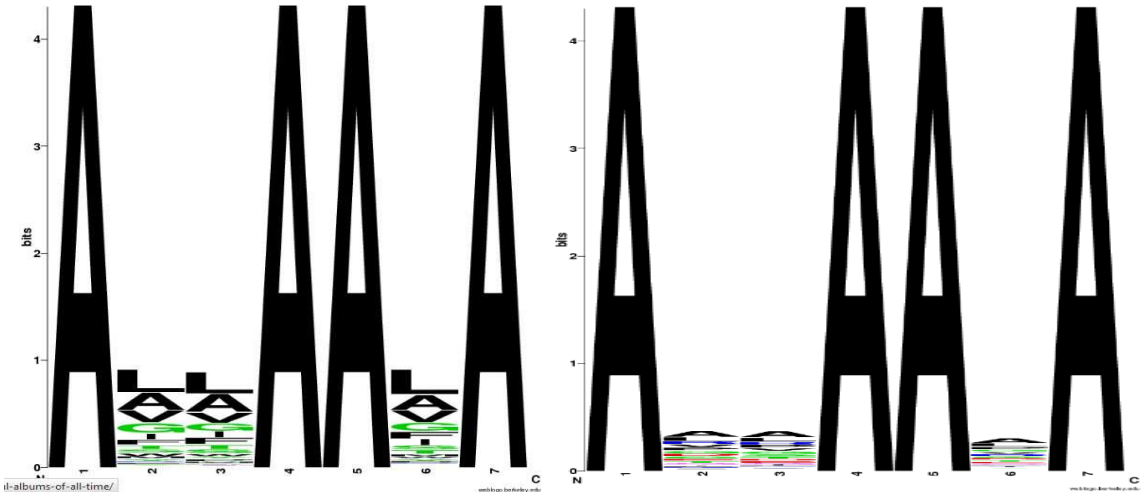

Abbildung 14 WebLogo für AA3AA2 im TM – Bereich und im nTM – Bereich

In der Abbildung 14 sind die WebLogos für die Paarung AA3AA2 im TM – Bereich und im nTM – Bereich dargestellt.

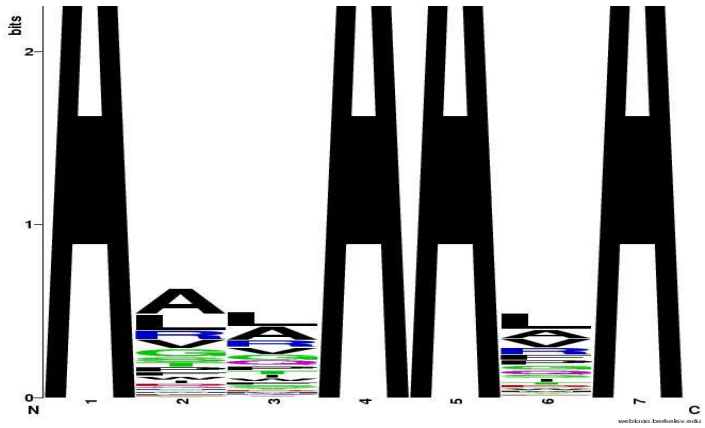

Abbildung 15 WebLogo für AA3AA2 im Transitionbereich

Das WebLogo zeigt die Ausprägungen des AA3AA2 Musterpaars im Transitionbereich.

In Abbildung 14 ist dies eindeutig zu erkennen, während im transmembranen Bereich auf den Zwischenpositionen vor allem Leucin, Alanin und Glycin vorkommen, also vor allem hydrophobe Aminosäuren, was logisch nachvollziehbar ist, da das gesamte Innere der Membran insgesamt einen hydrophoben Charakter besitzt, ist es im nichttransmembranen Bereich neben Alanin und Leucin vor allem Arginin. Im Transitionbereich (vergleiche Abbildung 15) wiederum herrscht eher ein Gemisch vor, was daher kommt, dass die Übergänge mal mehr im transmembranen Bereich und mal mehr im nichttransmembranen Bereich liegen und dann dementsprechend auch unterschiedliche Eigenschaften und Aminosäuren beinhalten.

Global über alle Familien gesehen ist es jedoch kaum der Fall, dass eine Paarung überhaupt nicht vorkommt oder nur in einem Bereich, lediglich die Häufigkeit ist anders.

4.3 Normierung der Paarungen

Durch eine Normierung für die 50 häufigsten Muster in transmembranen und nichttransmembranen Breichen wurde zwar eine leicht geänderte Reihenfolge erreicht (siehe Tabelle 7) , jedoch konnte nichts spezifisches festgestellt werden, global gesehen sind keine Paarungen auszumachen, die auffallend oft miteinander auftreten, da jede Paarung auftritt und anscheinend ohne Bedeutung, zumindest bei den jeweils 50 häufigsten.

Tabelle 7 Häufigste Muster im TM – und nTM -Bereich nach Normierung

nTM		TM	
Paarung	% der theoretisch möglichen Paarungen	Paarung	% der theoretisch möglichen Paarungen
AA3AA7	0,11514	AA2AA3	0,12456
AA2AA3	0,11138	AA2AA7	0,12262
AA2AA7	0,11306	AA2AG5	0,11112
AA3GA7	0,10493	AA3AG5	0,10727
AA7GA7	0,10324	AA2GA4	0,10716
AA7GA4	0,10305	AA3GA4	0,10360
AA3AG5	0,10257	AA2AG4	0,10213
AA3AG7	0,10251	AA2AL6	0,10101
AA2AG7	0,10239	AA3AG4	0,09914
AA2GA7	0,10174	AA3AL6	0,09950

Liste der zehn am häufigsten auftretenden Muster in den nTM – und Tm - Bereichen nach Normierung, in der ersten Spalte steht jeweils die gefundene Paarung, in der zweiten Spalte ist angegeben, wie viel Prozent der theoretisch möglichen Paarungen wirklich eingetreten sind.

Global über alle Familien betrachtet sticht also keine Paarung sonderlich hervor, betrachtet man jedoch die einzelnen Familien, ist zu erkennen, dass bestimmte Paarungen nur in diesen gehäuft vorkommen (siehe Tabelle 8).

Tabelle 8 Relative Häufigkeit von Musterpaarungen

Liste von Paarungen in nTM – und TM - Bereichen, die nur in der angegebenen Familie gehäuft auftreten, in der ersten Spalte steht jeweils die Proteinfamilie, in der zweiten die Paarung und in der dritten Spalte ist die relative Häufigkeit angegeben.

Relative Häufigkeit von Paarungen in einzelnen Familien					
nTM			TM		
Familie	Paarung	rel. HF in %	Familie	Paarung	rel. HF in %
PF09767	AA3SA3	15,46	PF09767	GA4GL1	4,84
PF09874	GL2IL4	21,21	PF09874	GV2VL4	16,78
PF09877	LY6PL1	22,22	PF09877	GL3VL4	12,36
PF09879	LG5SL2	23,21	PF09878	GL3IL4	11,43
PF09980	AA7SA6	13,18	PF09973	LF8SL2	9,87

Diese Paarungen scheinen also eine Ähnlichkeitsbeziehung darzustellen, jedoch kommt es auch hier nie vor, dass ein Muster X mit seinem Vorhandensein immer das Auftreten des Musters Z bewirkt.

In den Untersuchungen der Transition - Bereiche konnte für die einzelnen Familien nachgewiesen werden, dass häufig Paarungen mit einem langen Muster und einem kurzen auftreten, wie LF10VL4 in der Familie PF09877, oder IL4LF8 und IL4LF9 in der Familie PF09971. Diese beiden Paarungen kommen in der jeweiligen Familie zudem nicht im nichttransmembranen Bereich vor, wohl aber im transmembranen Bereich. Das häufige Auftreten langer Muster in diesem Bereich lässt darauf schließen, dass sie aufgrund ihrer Größe mitverantwortlich für die Stabilität der Proteinstruktur sind, zudem beginnen hier die Transporte in die Zelle, wofür konservierte Bereiche wichtig sind. Dafür sprechen auch WebLogos, wie in der Abbildung 16 zu erkennen.

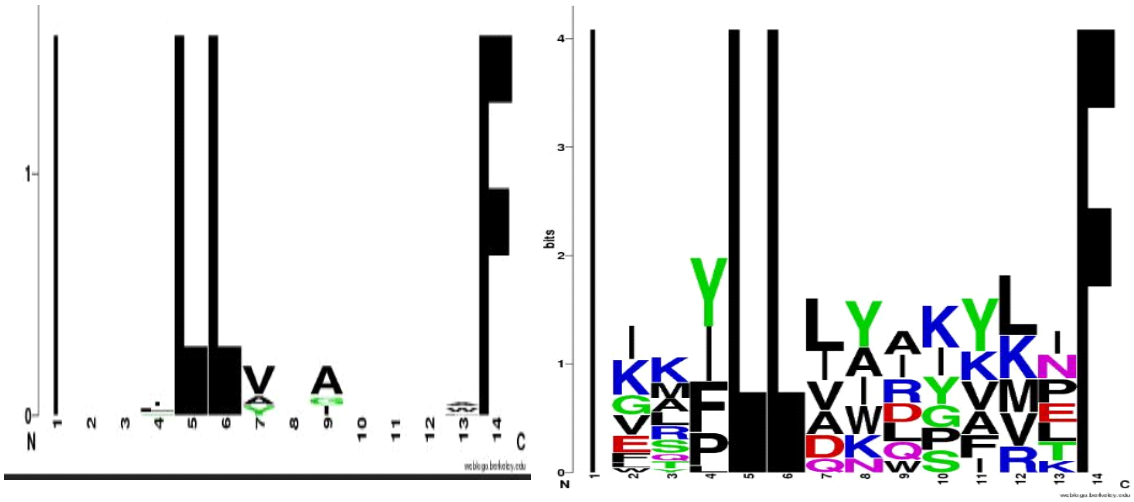

Abbildung 16 WebLogos für IL4LF8 im transmembranen Bereich und im Transition - Bereich

Dargestellt sind WebLogos für die Paarung IL4LF8 im TM - Bereich und im Transition – Bereich.

Während die Paarung IL4LF8 in der Transmembran zum einen selten und zum anderen ohne konservierte Zwischenpositionen vorliegt, erkennt man im Transition – Bereich, welche Aminosäuren vor allem auf welcher Zwischenposition zu finden sind. Jedoch ist es auch hier, wie bei allen angefertigten WebLogos für häufig auftretende Paarungen, dass es nie dazu kommt, dass wirklich nur eine Aminosäure konserviert an einer Zwischenposition auftritt, es sind stets mindestens zwei Aminosäuren auf einer Position möglich. Aufgrund dieser Tatsache lassen sich auch keine wirklich genauen regulären Ausdrücke für Muster(paarungen) in bestimmten Bereichen formulieren, die Zwischenpositionen sind einfach zu variabel. Anzumerken ist jedoch, dass auch nicht für alle Paarungen WebLogos erstellt werden konnten, was durch die bloße Menge an Paarungen (1225 für jeweils jeden Bereich und nochmals jeweils 1225 x 32 für die einzelnen Familien) verständlich ist.

Zusammenfassend ist zu sagen, dass in verschiedenen Bereichen von Transmembranproteinen auch verschiedene Musterpaarungen auftreten, welche sich auch in ihren Ausprägungen unterscheiden. Festgehalten werden muss auch, dass die am häufigsten auftretenden Musterpaarungen stets von den gleichen Mustern gebildet werden. Dabei handelt es sich vor allem um die Muster AA2, AA3, AL6 und GL1. Wie lässt sich dies erklären? Zum einen handelt es sich dabei um Muster, die auch allein zu den am häufigsten auftretenden Mustern in Transmembranproteinen gehören, zum anderen setzen sich zumindest ihre Start – und Endpositionen vorwiegend aus Alanin und Leucin zusammen, welche einfach die am häufigsten vorkommenden Aminosäuren sind, es ist also nur eine Frage der Wahrscheinlichkeit, dass diese Muster in so vielen Paarungen auftreten. Jedoch sollte man nicht außer Acht lassen, dass gerade diese oft vorkommenden Mustern eine prägende Funktion haben können, sei es für die Struktur des Proteins oder dessen Funktion.

Auch die verwendeten Daten, sprich die 32 Pfam Familien, beeinflussten selbstverständlich die Ergebnisse und bei genauerer Betrachtung dieser fällt auf, dass einige Familien nicht sehr aussagekräftig in Bezug auf Transmembran – Bereiche waren, da sie oft aus Proteinen bestehen, deren Sequenzen nur outside - Bereiche besitzen. Somit machte es auch nur bedingt Sinn, die Paarungen in den einzelnen Familien zu untersuchen, besser ist die globale Betrachtung für alle Proteinfamilien zusammen bzw. wäre es möglicherweise für weitere Untersuchungen von Vorteil, die Pfam Familien zu filtern, damit diese Familien keinen zu großen Einfluss auf mögliche Ergebnisse nehmen. Es konnte festgestellt werden, welche Paarungen am häufigsten in allen Proteinsequenzen der Familien auftreten und welche nur in bestimmen Familien gehäuft auftreten.

Ob diese Paarungen nun strukturunterstützend wirken oder für die Funktion des Proteins von Nöten sind, kann nur geklärt werden, wenn man weiß, ob die einzelnen Muster einer bestimmten Funktion zugeordnet werden können. Aus diesem Grund wurde im folgenden Schritt versucht, einzelnen Mustern eben solch eine mögliche Funktion zuzuordnen, da nur an Hand der vorliegenden Häufigkeitstabellen keine Aussagen darüber getroffen werden können, welchen Einfluss die Musterpaarungen im Protein besitzen.

4.4 Mögliche Funktionen der Muster

4.4.1 Mögliche Funktionen der einzelnen Muster

Nach einer intensiven Internetrecherche in Betreff auf Motive und beschriebene Funktionen in der Literatur und der Suche nach geeigneten Mustern/Pattern auf den verschiedenen Datenbanken, in denen die Muster von Gerstein et al. Gefunden werden könnten, wurden diese in einer Tabelle zusammengefasst mit passendem regulärem Ausdruck, dem Fundort und der Funktionsbeschreibung. In der Literatur wurden vor allem Muster gefunden, welche aus drei bis acht Positionen bestehen, weshalb zum einen untersucht wurde, ob Gersteinmuster in diesen vorkommen und zum anderen, ob die Muster aus der Literatur ein kleinerer Teil der Gersteinmuster sind. Dies war der erste Schritt, um mögliche Funktionen der 50 verwendeten Muster dieser Arbeit herauszufinden und dies in Einklang mit der Co - Co – Occurence zu bringen, damit Ähnlichkeitsbeziehungen zwischen diesen wohldefinierten Mustern möglicherweise erklärt werden können.

Für die Muster aus der Literatur wurde festgestellt, dass viele in den Ausprägungen der Gersteinmuster in den 32 verwendeten Pfam Familien auftauchen, was vor allem durch ihre geringe Länge begründet werden kann. Jedoch ließen sich einzelne Literaturmuster finden, die nur in den Ausprägungen von ein bis drei Gersteinmustern vorkommen, sie sind in der Tabelle 9 aufgelistet, ein X steht für eine beliebige Aminosäure.

Tabelle 9 Muster aus Literatur in Gersteinmustern

Abgebildet sind Muster aus der Literatur, welche höchstens in drei verschiedenen Gersteinmustern gefunden wurden.

Muster aus Literatur	gefunden in Gersteinmuster
[GAS]XXXGXXXG	LF10 PG9 PG10
CGPG	AG5 PG10 VF8
FXXXXXXLL	LF10 PG10
GXXXGXXX[GST]	LF10 PG9 PG10
GDAY	GG4 GS5 GY8
LFNEF	LF9
DPPY	AG7

Man erkennt, dass Muster, die wenig konservierte Positionen besitzen, auch häufiger gefunden werden, vor allem in großen Gersteinmustern wie LF10 oder PG10. Daher scheinen diese auch ungeeignet dazu, Gersteinmustern mögliche Funktionen zuzuordnen. Hinzu kommt, dass die Literaturmuster in Tabelle 9 meist Funktionen haben, die sich nicht auf die Transmembran beziehen, sie sogar oft nicht in dieser liegen.

Eine Ausnahme stellt dabei das Muster GXXXGXXX[GST] dar, welches in der Literatur als Glycinzipper in der Transmembran beschrieben wird. Gefunden wurde es in Ausprägungen von LF10, PG9 und PG10, was vermuten lässt, dass diese von Gerstein beschriebenen Muster auch als Glycinzipper dienen könnten, sie also einen Zipper enthalten und durch zusätzliche Aminosäuren diesen einfassen, um somit möglicherweise die Struktur des Proteins zu festigen. Ermittelt man nun, welche Gersteinmuster in den verschiedenen Literaturmustern auftreten, so fällt als erstes auf, dass das GG4 Muster unter anderm im Muster GXXXGXXX[GST] vorkommt, es lässt sich also aussagen, dass GG4 als Glycinzipper wirkt, was jedoch auch schon Mark Gerstein festgestellt hatte.

Weitere gefundene Gersteinmuster sind in der folgenden Tabelle 10 zusammengefasst:

Tabelle 10 Gersteinmuster in Literaturmustern

Dargestellt sind Gersteinmuster, die in Mustern aus der Literatur gefunden wurden.

Gersteinmuster	gefunden in Muster aus Literatur
AA2	EGYATA
GA4	EGYATA
AG4	[GAS]XXXGXXXG
AG5	[GA]XXXXGK[ST]
GG4	LIXXGVXXGVXXT [GAS]XXXGXXXG GXXXGXXX[GST]
GV1	LIXXGVXXGVXXT
IV4	LIXXGVXXGVXXT
GS4	GXXXGXXX[GST]
GS5	GXXRXSG
PG10	ISXPXXFXHXXHVG
SL2	[ST]X[IVL]
SG4	IYTSGTTGXPK [GAS]XXXGXXXG

Auffallend ist, dass vor allem Muster in der Tabelle auftauchen, welche ein Glycin als Start- oder Endposition besitzen, diese Muster scheinen also verstärkt Einfluss auf Funktion und Struktur eines Proteins zu besitzen. Dies wird auch deutlich, wenn man die Funktionen und Beschreibungen der Literaturmuster betrachtet. [GAS]XXXGXXXG und GXXXGXXX[GST] sind Glycinzipper in der Transmembran, da GG4 in diesen gefunden wurde, wird wieder deutlich, dass es eine ähnliche Funktion innehaben muss. Aber auch die Muster AG4, GS4 und SG4 wurden in diesen Literaturmustern gefunden, jedoch nicht konserviert, wie es bei GG4 der Fall ist. Aus diesem Grund werden diese Treffer im Folgenden nicht weiter beachtet, da es höchst spekulative Vemutungen wären, wenn man einem Gersteinmuster, welches auf variablen Stellen eines größeren Musters gefunden wurde, dessen Funktion zuordnen möchte.

Auf die anderen Gersteinmuster aus Tabelle 10, die konserviert vorkommen, wird im späteren Verlauf der Arbeit genauer eingegangen. Zunächst soll beschrieben werden, wie Gersteinmuster in den Mustern/Motiven/Pattern der durchsuchten Datenbanken MeMotif, ELM und Prosite gesucht wurden. Hierfür wurden die regulären Ausdrücke der Datenbanken mit Hilfe des Perl – Scripts von Brad Bowman in Strings umgewandelt und die 50 Gersteinmuster in diesen gesucht. Da es zu einer sehr hohen Trefferanzahl kam, wurden die Treffer zunächst dahingehend untersucht, ob die Gersteinmuster wirklich konserviert gefunden wurden. Durch diese Filterung der Treffer konnte eine Liste von 26 Gersteinmustern erstellt werden, die in Motiven/Pattern aus den Datenbanken konserviert sind (siehe Auszug in Tabelle 11 und Anhang 2).

Tabelle 11 Gersteinmuster in Datenbankmustern

Gersteinmuster	regulärer Ausdruck der Datenbanken	Funktion des Musters in Datenbanken
AA2	EGYATA	part of the active center for primase function
AA3	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
AG4	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L L-x-A-A-x(2,3)-A-P-L-[AT]-G-I A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G A-[GL]-[LM]-[AS]-A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I	Reentrant loop in chloride channels Reentrant loop in chloride channels protein structural stability and dimerization reentrant loop in chloride channels
AL6	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L L-x-A-A-x(2,3)-A-P-L-[AT]-G-I A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	Reentrant loop in chloride channels Reentrant loop in chloride channels protein structural stability and dimerization
AG7	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S T-A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S A-[LMF]-x-[GAT]-T-[LIVMF]-x-G-x-[LIVMF]-x(7)-P	electron transport chain part of the binding pocket of heme b and the close inhibitor stigmatellin electron transport chain This protein probably forms a transmembrane proton channel
FG1	F-F-G-V-x(2,3)-F F-F-G-V-x-[AGT]-[FIM]	Membrane-water interface motif interactions with cofactors, lipid-exposure motif
GL2	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L V-G-x-L-[ACTV]-[AGT]-[IL]-[AGV] V-x(0,2)-G-x(3)-G-x(1,2)-L	Reentrant loop in chloride channels involved in helix-helix packing -
GV2	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
GL3	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L G-V-x(1,2)-L-[ILV]-[GL]-[TV]-[LV]-[LM] G-[CGS]-[AG]-L-[FL]-x-A-x-H-G-[AS]-x-[IV]	Reentrant loop in chloride channels putative helix-helix interaction nvolved in the binding of the mononuclase of M chain
GA4	L-G-[ILM]-G-x-H-x(1,3)-A-x(0,2)-F Q-I-[LV]-T-G-[IL]-[FLV]-[LM]-A-M-H-Y EGYATA	conserved site for binding of cardiolipin involved in heme binding part of the active center for primase function
GG4	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S G-L-Y-x-G G-L-Y-Y-G-S-Y	electron transport chain iron ion binding, electron transport Structured-loop motif connecting two helices

In der Tabelle 11 sind Gersteinmuster zusammengefasst, welche in Datenbankmustern und auch Literaturmustern konserviert gefunden wurden. In der ersten Spalte steht das gefundene Gersteinmuster, in der zweiten der reguläre Ausdruck des Datenbankmusters und in der dritten Zeile steht die Funktion des Datenbankmusters.

Um zu zeigen, aus welchen Quellen die regulären Ausdrücke stammen, wurde das Diagramm in der Abbildung 17 erstellt.

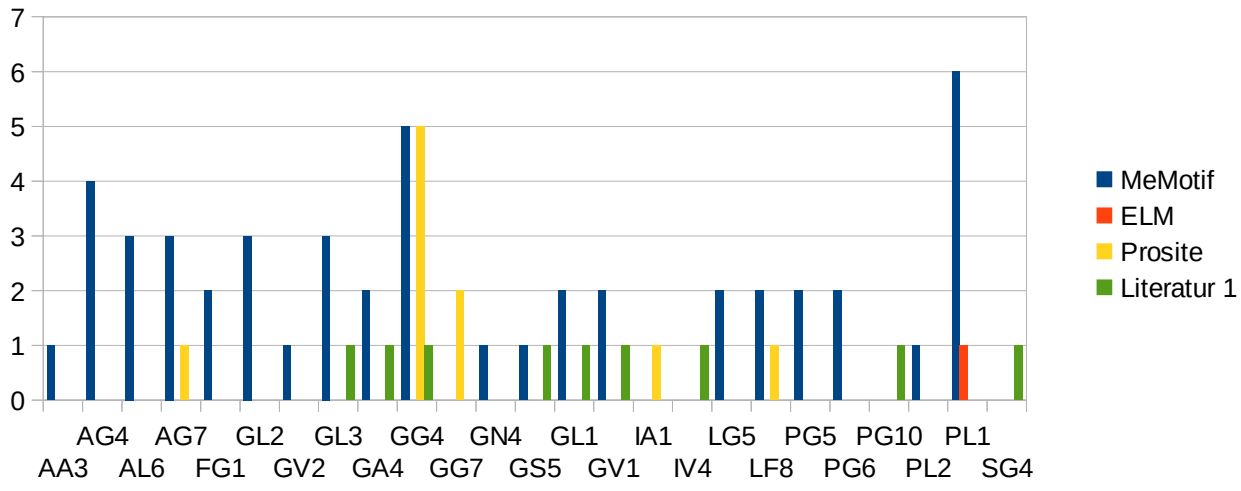

Abbildung 17 Diagramm über Fundorte der regulären Ausdrücke

Das Diagramm zeigt für die konserviert vorgefundenen Gersteinmuster, von welcher Datenbank die durchsuchten Muster/Pattern stammen oder ob sie in Mustern aus der Literatur gefunden wurden und wie oft sie dort gefunden wurden.

Es ist zu erkennen, dass die Gersteinmuster vor allem in Mustern von MeMotif gefunden wurden, am häufigsten war das GG4 Muster konserviert enthalten. Dieses Muster scheint also eine wichtige Funktion innerhalb der Transmembran innezuhaben. Zudem zeigt diese Statistik, dass ELMs wenig geeignet dafür scheinen, auf Enthalten kleinerer konservierter Muster hin untersucht zu werden bzw. dass einfach so gut wie keine Gersteinmuster in diesen linearen Mustern vorkommen. Zudem wurde etwa die Hälfte der gefundenen Gersteinmuster nur einmal oder zweimal in den Datenbankmustern entdeckt. Dies lässt wiederum darauf schließen, dass die Gersteinmuster allgemein nicht sehr oft in größeren konserviert vorliegen, was nicht unbedingt den Erwartungen entspricht, da diese Muster ja laut Gerstein et al. besonders oft in der Transmembran vorkommen und daher von ihm beschrieben wurden. Deshalb kann es auch nicht verwundern, dass sich die meisten Treffer in den Mustern von MeMotif fanden. Diese Datenbank beinhaltet schließlich nur Muster aus Membranproteinen, während Prosite und ELM diese Unterscheidung nicht extra treffen. Es scheint daher ratsam, sich ausgiebiger mit der Memotif Datenbank auseinanderzusetzen, wenn man sich in seinen Forschungen mit Membranproteinen beschäftigt, es liegen zwar lediglich 53 annotierte Sequenzmuster vor, jedoch beinhalten schon diese sehr oft die Gersteinmuster.

Anhand dieser Ergebnisse kann nun versucht werden, bestimmten Gersteinmustern eine mögliche Funktion auf Grundlage der Funktionen der Datenbankmustern zuzuordnen. Zusätzlich wurden kleinere Datenbankmuster in Gersteinmustern gesucht, auch hier kristallisierten sich einzelne Muster heraus, die nur in einem Gersteinmuster gefunden wurden.

Dabei handelt es sich um das [DE][DES][DEGAS]F[SGAD][DEAP][LVIMFD] Muster von der ELM, welches im PG10 Muster gefunden wurde und mit Domänen von Adapterproteinen interagieren soll und um zwei Muster von MeMotif. Dies ist zum einen das V[ST][ILM[AT]TV Muster, welches ein Filtermotiv in Kaliumkanälen darstellt und in Ausprägungen des VG6 Musters gefunden wurde und zum anderen das VGXL[ACTV][AGT][IL][AGV] Muster, welches in Ausprägungen des VF8 Musters gefunden wurde und in dem Helix – Helix Packing involviert ist. Während das [DE][DES][DEGAS]F[SGAD][DEAP][LVIMFD] Muster nur eine konservierte Position beinhaltet, sind es bei den beiden MeMotif Mustern immerhin jeweils vier. Dies ist auch eine Feststellung, die im Laufe der Arbeit getroffen werden konnte. Insgesamt gesehen eignen sich vor allem Muster von MeMotif für die Suche, da die Pattern/Muster von ELM und Prosite oft nur wenige konservierte Positionen beinhalten. Wie dem auch sei, könnte man anhand des Vorkommens in den VG6 bzw. VF8 Mustern daruf schließen, dass eben diese Gersteinmuster ähnliche oder gleiche Funktionen innehaben, wie die Muster von

MeMotif. Somit ist es anhand der bisherigen Ergebnisse theoretisch möglich, 28 der 50 Gersteinmustern eine Funktion zuzuordnen, da einige Funktionen jedoch nicht genügend Aussagekraft in Bezug auf Transmembranproteine besitzen, wie z. Bsp. das Muster EGYATA, welches lediglich Teil des aktiven Zentrums ist, welches für die Primasefunktion zuständig ist, scheint es angebracht, die Tabelle noch zu filtern. Nach diesem Schritt verbleiben 19 Gersteinmuster, denen möglicherweise eine Funktion betreffs des Vorkommens in Transmembranproteinen zugeordnet werden kann, sie sind in der Tabelle 12 zusammengefasst.

Tabelle 12 Zuordnung möglicher Funktionen

In der Tabelle 12 sind Gersteinmuster zusammengefasst, welche in Datenbankmustern und auch Literaturmustern konserviert gefunden wurden und deren Funktion sich auf die Transmembran bezieht. In der ersten Spalte steht das gefundene Gersteinmuster, in der zweiten der reguläre Ausdruck des Datenbankmusters und in der dritten Zeile steht die Funktion des Datenbankmusters.

Gersteinmuster	regulärer Ausdruck der Datenbanken	Funktion des Musters in Datenbanken
AA3	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
AG4	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	protein structural stability and dimerization
	A-[GL]-[LM]-[AS]-A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I	reentrant loop in chloride channels
AL6	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	protein structural stability and dimerization
AG7	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	A-[LMF]-x-[GAT]-T-[LIVMF]-x-G-x-[LIVMF]-x(7)-P	probably forms a transmembrane proton channel
FG1	F-F-G-V-x(2,3)-F	Membrane-water interface motif
GL2	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	V-G-x-L-[ACTV]-[AGT]-[IL]-[AGV]	involved in helix-helix packing
GV2	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
GL3	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	G-V-x(1,2)-L-[ILV]-[GL]-[TV]-[LV]-[LM]	putative helix-helix interaction
GG4	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	G-L-Y-x-G	iron ion binding, electron transport
	G-L-Y-Y-G-S-Y	Structured-loop motif connecting two helices
	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	LIXXGVXXGVXXT	dimerization motif for transmembrane alpha-helices
	[GAS]XXXGXXXG	distinct preference for right-handed packing against a heterologous helix surface
GS5	G-L-Y-Y-G-S-Y	Structured-loop motif connecting two helices
GL1	G-L-Y-x-G	iron ion binding, electron transport
	G-L-Y-Y-G-S-Y	Structured-loop motif connecting two helices
GV1	G-V-x(1,2)-L-[ILV]-[GL]-[TV]-[LV]-[LM]	putative helix-helix interaction
	F-F-G-V-x(2,3)-F	Membrane-water interface motif
	LIXXGVXXGVXXT	dimerization motif for transmembrane alpha-helices
IV4	LIXXGVXXGVXXT	dimerization motif for transmembrane alpha-helices
LG5	L-x-[AIV]-x-[HW]-G-A-T	modulator of transmembrane protein function
PG5	P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]-[AGST]-[FILV]	interaction with different membrane lipids and thecytochrome subunit
PG6	L-W-x-[AGI]-Y-P-[FIV]-[ILV]-W-[IL]-[ILV]-G	Proline-kink motif
PL1	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
	P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]-[AGST]-[FILV]	iteraction with different membrane lipids and thecytochrome subunit
	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	protein structural stability and dimerization
	A-[GL]-[LM]-[AS]-A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I	reentrant loop in chloride channels
Datenbankmuster	gefunden in Gersteinmuster	Funktion des Musters in Datenbanken
V[ST][ILM[AT]TV	VG6	filter motiv in potassium channels
VGXL[ACTV][AGT][IL][AGV]	VF8	involved in helix-helix packing

Diese Gersteinmuster sind vor allem in Mustern aus MeMotif gefunden worden, welche sich in einem Reentrantloop in Chloridkanälen befinden, welche durch das Durchqueren der Membran höchstwahrscheinlich zur Stabilität der Struktur des gesamten Proteins beitragen. Andere häufige Funktionen sind Ionenkanäle oder Helix – Helix Interaktionen,

welche entscheidende Faktoren für die Proteinstruktur sind. Auffallend ist, dass vor allem für Gersteinmuster mit einer Länge von sieben oder weniger Aminosäuren Datenbankmuster gefunden wurden, was sich rein statistisch gesehen erklären ist. Gersteinmuster mit zwei oder drei Aminosäuren können einfach öfter gefunden werden, einige größere Gersteinmuster sind sogar länger als die Datenbankmuster, können also überhaupt nicht in ihnen gefunden werden, während kleinere Muster mit Aminosäuren als Start – und Endposition, welche auch noch in der Natur am häufigsten auftreten, sogar mehrmales in einem Datenbankmuster gefunden werden können. Daher müsste auch hier eine Normierung durchgeführt werden, um diejenigen Gersteinmuster auszuschließen, die nur einen geringen Anteil, bezogen auf die Anzahl der Positionen, des Gesamtmusters aus den Datenbanken, in dem sie gefunden wurden, ausmachen. Somit bleiben letztendlich nur sieben Gersteinmuster übrig, welche konserviert in Datenbankmustern gefunden wurden, die eine Funktion innehaben, welche mit der Transmembran in Verbindung gebracht werden kann und bei denen die Gersteinmuster zumindestest die Hälfte der Positionen des Datenbankmusters einnimmt. Diese Gersteinmuster lauten AG4, AL6, AG7, GG4, GS5, LG5 und PG6, die Funktionen der dazugehörigen Datenbankmuster sind der Tabelle 12 zu entnehmen. Von diesen sieben Mustern sind vor allem das GG4 und das GS5 Muster hervorzuheben, welche am besten mit den jeweiligen Datenbankmustern übereinstimmen. Bei GG4 ist dies das GLYXG Muster, welches für den Elektronentransport wichtig ist, beide Muster haben die gleiche Länge, also kann ausgesagt werden, dass die Ausprägung GLYXG des GG4 Musters womöglich immer diese Funktion innehat und sich deshalb entweder oft im Transitionbereich finden lässt oder am Anfang einer TM – Helix. Bei GS5 handelt es sich bei dem besten Treffer um das Muster GLYXGSY, das Datenbankmuster ist also lediglich um eine Position länger. Die Funktion des GLYYGSY Musters ist laut MeMotif, wo es als Eintrag gefunden wurde, dass es zwei Helices durch einen Loop miteinander verbindet, hier lässt sich für die Ausprägung GLYYGS des GS5 Musters aussagen, dass es mit großer Wahrscheinlichkeit öfters in solchen Loops auftritt. Die besten Treffer beziehen sich also zusammenfassend meist auf Muster, welche zwischen den Helices vorkommen und so für deren Stabilität sorgen bzw. in diesem Bereich die Struktur des Proteins bestimmt wird. Der Transitionbereich ist auch generell von großer Wichtigkeit, da hier die Stoffaustausche durch die Membran stattfinden bzw. beginnen. Es ist auch noch festzustellen, dass nur sehr wenige Gersteinmuster wirklich gut mit den Datenbankmustern übereinstimmen, meist sind sie nur ein kleiner Teil der großen Datenbankmuster, weshalb es als zu unsicher angesehen wird, Aussagen über mögliche Funktionen für viele Gersteinmuster zu treffen.

Da nun Rückschlüsse auf Funktionen einiger Muster gezogen werden konnten, besteht der abschließende Schritt dieser Arbeit im Folgenden nun darin, die Funktionen einzelner Muster mit der Co - Co – Occurence zu korrelieren, um Aussagen darüber treffen zu können, weshalb bestimmte Paarungen gehäuft auftreten und ob das Auftreten dieser Paarungen möglicherweise das Vorhandensein weiterer Muster in der Proteinsequenz mit sich führt. Dabei wiederum kann man versuchen auch Gersteinmustern, die nicht so gut mit den Datenbankmustern, in denen sie gefunden wurden, übereinstimmen, eine mögliche Funktion zuzuordnen, da nun immer zwei Muster im Kontext zueinander betrachtet werden.

4.4.2 Rückschlüsse auf Funktionen in Hinblick auf die Co – Co – Occurence

Bei der Betrachtung von zwei Mustern, welche zusammen in einer Proteinsequenz vorkommen, sollte es möglich sein, anhand der bisherigen Ergebnisse Rückschlüsse darüber ziehen zu können, welche Funktionen diese innehaben, wenn diese Paarung gehäuft auftritt, um somit eine Erklärung für ihr Auftreten zu erhalten und zusätzlich möglicherweise klären zu können, welche andere Muster im Zusammenhang mit den Mustern der jeweiligen Paarung wirken bzw. die Struktur des Proteins gemeinsam festigen oder erst bilden.

Zunächst sollen die jeweils am häufigsten auftretenden Paarungen in den transmembranen und den Transitionbereichen daraufhin untersucht werden, die Übersicht über diese stellt die Tabelle 5 dar. Für die Paarungen der nichttransmembranen Bereiche können keine Aussagen getroffen werden, da keine Muster gefunden wurden, die explizit in diesen Bereichen vorkommen bzw. dort eine Funktion innehaben, zudem sind die nichttransmembranen Bereiche nicht entscheidend für die Struktur der Transmembranproteine, die hautpsächlich durch die Helices im TM – Bereich und die Loops im Transitionbereich beeinflusst wird.

Das jeweils am häufigsten in den einzelnen Bereichen auftretende Paar ist AA2AA3, jedoch konnte dem AA2 Muster im Verlauf dieser Arbeit keine Funktion zugeordnet werden, bei AA3 könnte es sich immerhin um einen häufigen Bestandteil eines Reentrantloops handeln. Für Transitionbereiche könnte man also vermuten, dass auch AA2 häufig in solchen zu finden ist und die beiden Muster so zusammenwirken, dass sie die Stabilität der Proteinstruktur maßgeblich beeinflussen. Im Weiteren sollen jedoch Paarungen, welche das AA2 Muster oder andere Muster enthalten, von den Untersuchungen ausgenommen werden, damit die getroffenen Vermutungen und

Aussagen mit einer höheren Wahrscheinlichkeit auch der Wahrheit entsprechen. Die nächsthäufigsten Paarungen im Transitionbereich, deren einzelne Muster konserviert in Datenbankmustern mit einer Funktion gefunden wurden, sind AA3GL1, AA3AG4, GL1GL3, GL1GL2, AA3AL6 und AL6GL1. Die Muster GL1, AG4, GL3, GL2 und ebenso AL6 wurden in Datenbankmustern von MeMotif gefunden, welche ihre Funktion als Bestandteil eines Reentrantloops innehaben. Alle diese Muster scheinen in diesen Loops eine struktur- und stabilitätsgebende Funktion zu besitzen, weshalb es logisch erscheint, dass Paarungen, in den die Muster aus den Arbeiten von Gerstein et al. Enthalten sind, auch am häufigsten in Transitionbereichen vorkommen. Erwähnenswert ist in diesem Zusammenhang vor allem das MeMotif Muster A[AS][FIV][NR]APL[AT]GIXL bzw. A[AS][FIV][NR]APL[AT]GIL, in dem AL6 und GL3 bzw. GL2 lediglich durch eine Aminosäure voneinander getrennt, hintereinander auftreten. Somit wird immer mehr ersichtlich, wie diese Muster zusammen die Proteinstruktur prägen, vor allem, wenn sie nebeneinander in der Sequenz auftauchen. Durch diesen Fund stellt sich jedoch auch die Frage, wie Gersteinmuster zukünftig in Untersuchungen genutzt werden sollen, sprich, ob man schaut, wie sie zusammen Einfluss auf Struktur und Funktion des Proteins nehmen, wenn sie in unterschiedlichen Transitionbereichen oder auch Helices auftreten, oder ab man versucht, möglichst Aminosäureketten zu finden, in denen möglichst viele Gersteinmuster hintereinander, ohne großen Abstand zueinander auftauchen und so versucht, die Bedeutung dieser Gersteinmuster zu ergründen.

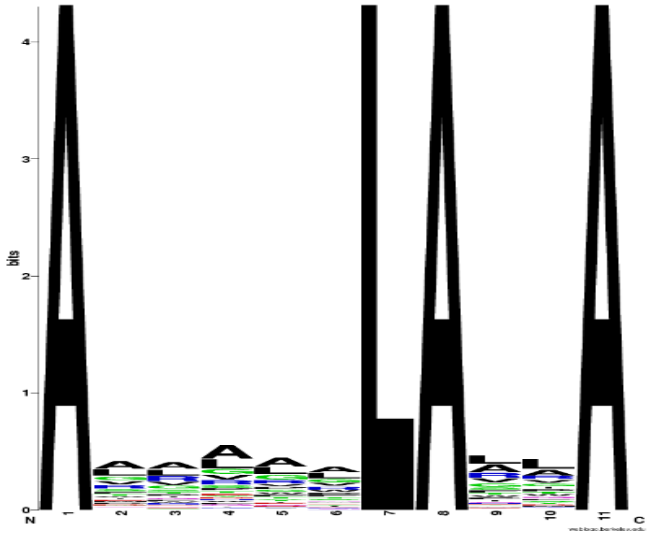

Abbildung 18 WebLogo für AL6AA3 im Transitionbereich

Zu sehen ist das WebLogo für die Paarung AL6AA3 im Transitionbereich.

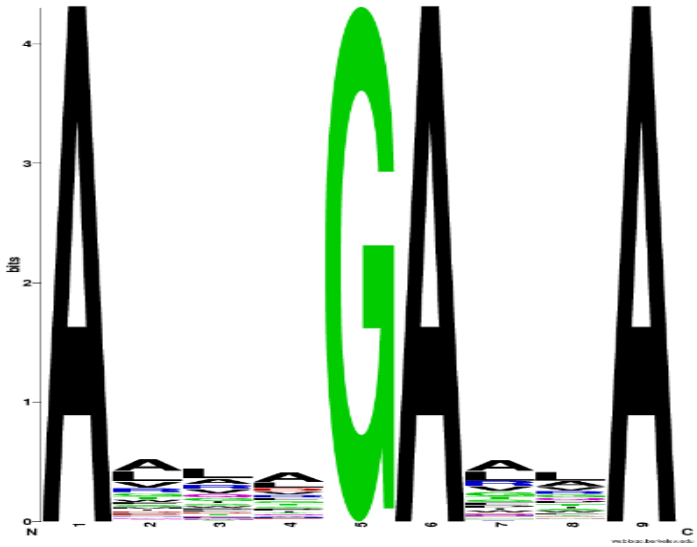

Abbildung 19 WebLogo für AG4AA3 im Transitionbereich

Abgebildet sind die Ausprägungen der Paarung AG4AA4 im Transitionbereich als WebLogo.

Während das WebLogo der Paarung AL6AA3, dargestellt in Abbildung 18, sich als nicht sonderlich auffällig für den Transitionbereich erweist, kann man bei der Paarung AG4AA3 (zu sehen in Abbildung 19) immerhin erkennen, dass auf der dritten Zwischenposition des AG4 Musters als dritthäufigste Aminosäure Asparaginsäure auftritt. Diese besitzt einen polaren und hydrophilen Charakter, ist also meist im nichttransmembranen Bereich vorhanden, was nahe legt, dass der AG4 Teil der Paarung in diesem Bereich liegt und das AA3 Muster den Transitionbereich ausmacht, was jedoch lediglich eine Vermutung ist. Sollte dies jedoch der Fall sein, würde dies bedeuten, dass Muster über bestimmte Bereiche hinaus Einfluss aufeinander nehmen könnten, sprich dass womöglich Muster im nichttransmembranen Bereich Einfluss auf die Struktur von Helicesbereiche nehmen. Deshalb ist es definitiv von Nöten, die einzelnen Bereiche eines Transmembranproteins im Zusammenhang miteinander zu betrachten, um wirklich klären zu können, wie bestimmte konservierte Bereich einer Proteinsequenz die Struktur und Funktion eines Proteins bestimmen.

Für den transmembranen Bereich sind Muster, welche in Reentrantloops vorkommen, nicht verwendbar, da es hier um die Helicesbereiche geht. Somit fallen jedoch auch einige Gersteinmuster, denen in Bezug auf die Transitionbereiche eine mögliche Funktion zugeordnet werden konnte, aus den zu untersuchenden Mustern heraus, weshalb nicht die allerhäufigsten Musterpaarungen im transmembranen Bereich mit einer Funktion in Einklang gebracht werden konnten. Zudem wurden nur die Paarungen betrachtet, bei denen die einzelnen Partnermuster in jeweils unterschiedlichen Helices

auftraten, um so zum einen zu vermeiden, dass Muster ineinander gefunden werden und zum anderen sollte so versucht werden, Muster zu finden, die für den Zusammenhalt und Interaktionen zwischen den Helices verantwortlich sind. Die häufigsten Musterpaarungen im TM – Bereich sind dabei in der Tabelle 13 abgebildet.

Tabelle 13 Häufigste Paarungen TM – TM

Abgebildet sind die häufigsten Paarungen im transmembranen Bereich zwischen Mustern in unterschiedlichen Helices und ihre Anzahl.

Paarung	Anzahl
AA3-AA2	33967
AL6-AA2	31343
AL6-AA3	30168
GL3-AA2	26975
AL6-GL3	26746
AL6-VL4	26691
AA2-VL4	26325
GL3-AA3	25880
AA3-VL4	25470
GL3-VL4	24072

Letztendlich konnten so die folgenden Paarungen herausgefiltert werden, welche noch relativ häufig im transmembranen Bereich auftreten und denen möglicherweise eine Funktion in diesem zugeordnet werden kann: GL2GL3, GL3LG5 und GL2LG5. Das GL2 Muster wurde im MeMotif Muster VGXL[ACTV][AGT][IL][AVG] gefunden, welches in das Helix – Helix Packing benachbarter Helices involviert ist, was damit zusammenhängt, dass benachbarte Helices bevorzugt über Segmente miteinander interagieren, die mindestens 10 Reste lang sind, zum anderen lässt es sich mit den enthaltenen kleinen Resten wie Glycin, Alanin und Threonin erklären, die oft in Helix – Helix Interaktionen involviert sind[12].

Das Gersteinmuster GL3 wiederum wurde im MeMotif Muster GVXL[ILV][GL][TV][LV][LM] gefunden, welches jedoch auch ein mutmaßliches Helix – Helix Interaktionsmuster darstellt.

Das Gersteinmuster LG5 wurde im Muster FXXXLXXL[IL]G und im Muster LX[AIV]X[HW]GAT gefunden, beide stammen ebenfalls von der MeMotif Datenbank. FXXXLXXL[IL]G scheint durch das Phenylalanin in Interaktionen mit den Lipiden der Doppellipidschicht involviert zu sein und ansonsten rein struktureller Natur und das

Muster LX[AIV]X[HW]GAT findet sich an der Bindungsstelle verschiedener Lipide, beiden Mustern kann also eine ähnliche Aufgabe zugeteilt werden und somit könnte man auch vermuten, dass LG5 oft am Anfang oder Ende einer Helix zu finden ist und mit den Lipiden der Doppellipidschicht interagiert.

Für die drei vorhins erwähnten Paarungen bedeutet dies nun, dass GL2GL3 wohl hauptsächlich daran beteiligt ist, dass benachbarte Helices miteinander interagieren, um so die Struktur des Proteins zu festigen, da beide Partnermuster in MeMotif Mustern gefunden wurden, die eine solche Funktion innehaben. Die Paarungen GL2LG5 und GL3LG5 könnten in der Beziehung strukturgebend sein, dass sie den Zusammenhalt zwischen Beginn einer Helix und dem Mittelsegment der benachbarten Helix sichern. Es lässt sich also schlussfolgern, dass die Muster GL2 und GL3 häufig inmitten einer Transmembranhelix auftauchen und womöglich dazu beitragen, dass benachbarte Helices miteinander agieren und sie zusätzlich oft mit LG5 interagieren könnten, welches sich mutmaßlich vor allem im Anfangsbereich einer Helix, nahe der Transitionbereiche, finden lässt und somit alle drei Muster zusammen prägend für die Stabilität der Helices im Transmembranbereich sind. Erweitert man die Co – Co – Occurence Untersuchungen zu einer Co – Co – Co - * - Untersuchung, so ließe sich dies genauer untersuchen und möglicherweise sogar bestätigen.

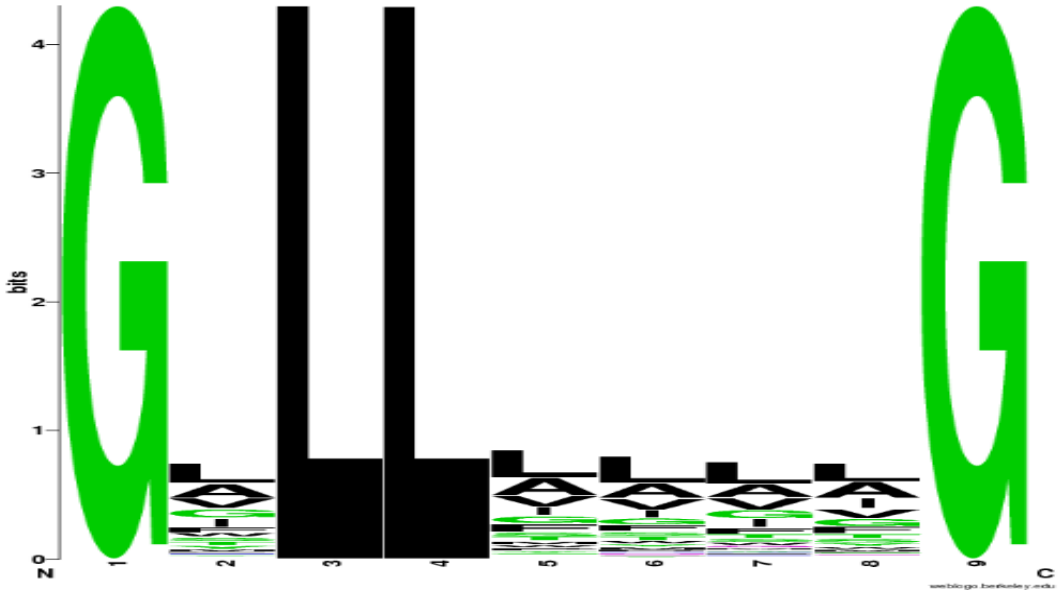

Abbildung 20 Paarung GL2LG5 im TM - Bereich

Abgebildet ist das WebLogo für die Paarung GL2LG5 im TM – Bereich.

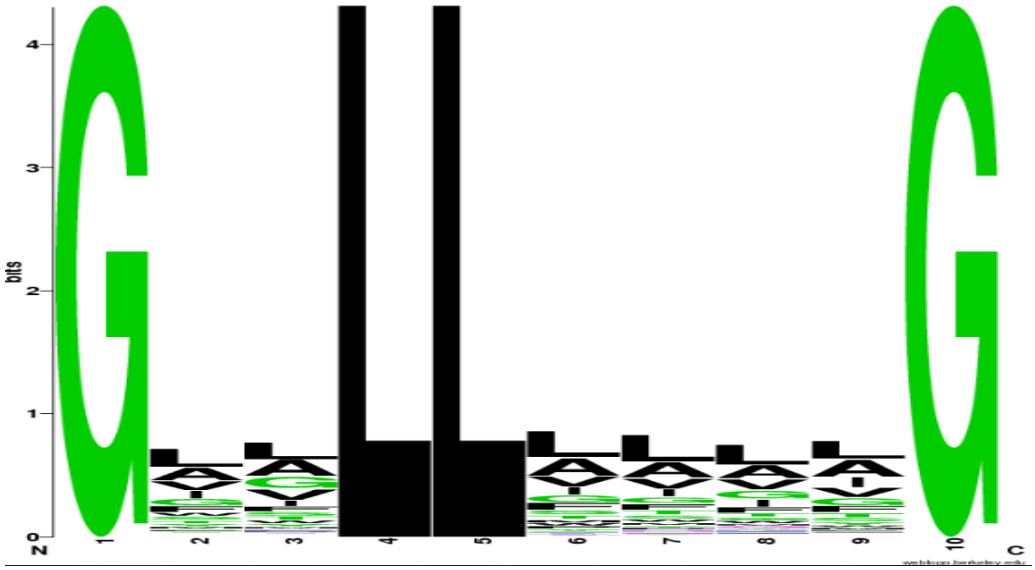

Abbildung 21 Paarung GL3LG5 im TM-Bereich

Abgebildet ist das WebLogo für die Paarung GL3LG5 im TM – Bereich.

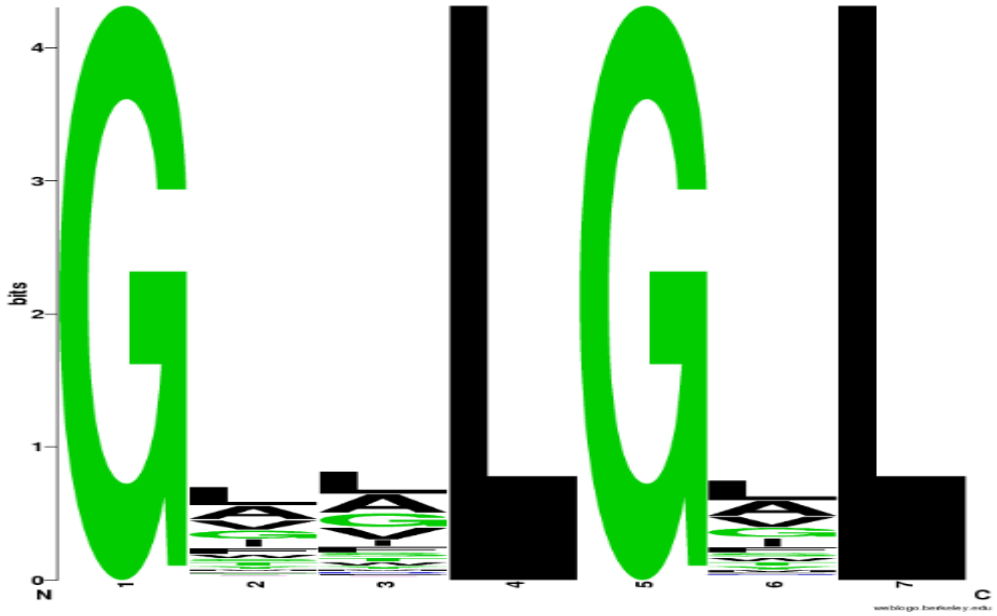

Abbildung 22 Paarung GL3GL2 im TM-Bereich

Abgebildet ist das WebLogo für die Paarung GL3GL2 im TM – Bereich.

Schaut man sich abschließend die WebLogos der drei Paarungen GL2LG5, GL3LG5 und GL3GL2 an (siehe Abbildungen 20, 21, 22), so unterstreichen diese zumindest, dass die Paarungen wohl oft in Helices auftauchen, da sie vor allem Leucin und Alanin auf den Zwischenpositionen beinhalten, welche hydrophob sind. Jedoch ist es nicht der Fall, dass auf einer Zwischenposition nur eine einzige Aminosäure konserviert vorkommt, was

möglicherweise anders wäre, würde man sich nur bestimmte gefundene Ausprägungen der Muster anschauen, sei es, dass man nur bestimme Familien betrachtet oder die Ausprägungen, die in bestimmten Helices auftreten.

Um die Ergebnisse der Funktionszuweisung einzelner Muster effektiv zu nutzen, wurde geschaut, welche dieser Gersteinmuster mit einer Funktion in Verbindung gebracht wurden, die unmittelbar mit der Struktur der Helices eines Proteins zusammenhängt, ungeachtet dessen, wie oft diese Muster in den 32 Proteinfamilien vorkommen bzw. wie häufig Paarungen auftreten, in denen diese enthalten sind. Dabei wurde auch darauf geachtet, kleinere Gersteinmuster, die in einem Datenbankmuster gefunden wurden, in dem auch ein größeres Gersteinmuster gefunden wurde, von der Untersuchung auszuklammern. Somit blieben neben dem VF8 Muster, in dem ein MeMotif Muster gefunden wurde, noch die Gersteinmuster GG4 und IV4 zu betrachten, die in dem Alpha – Helices dimerisierenden Muster LIXXGVXXGVXXT gefunden wurden. Betrachtet man die Häufigkeit der Paarungen dieser drei Muster im Transmembranbereich, so fällt auf,dass sie erstaunlicherweise relativ selten vorkommen, wenn man es mit den häufigsten Paarungen vergleicht. Während das Paar GG4VF8 nur 2123 mal auftritt, treten die Paarungen GG4IV4 und IV4VF8 immerhin 3173 bzw. 3295 mal auf, jedoch ist dies kein Vergleich zu der am häufigsten auftretenden Paarung AA2AA3 mit 33976 Treffern. Im Grunde genommen scheinen alle drei Muster dafür zu sorgen, dass sich die Helices dimerisieren bzw. die Helices in unmittelbarer Nähe zueinander richtig ,,gepackt“ werden, wenn man ihnen die Funktionen der Datenbankmuster zuordnen würde, was jedoch nicht mit Sicherheit gesagt werden kann, da es keine Argumente und Erkenntnisse gibt, die solche Vermutungen sicher untermauern.

Doch auch wenn man nicht exakt sagen kann, dass diese Gersteinmuster und deren Paarungen diese Funktionen innehaben, ist es doch relativ wahrscheinlich, da sie jeweils einen großen Anteil der Positionen der Datenbankmuster einnehmen bzw. bei VF8 der einzige Unterschied in der Aminosäurenabfolge derjenige ist, dass das Datenbankmuster VGXL[ACTV][AGT][IL][AGV] lediglich aus acht statt neun Aminosäuren besteht, das Phenylalanin an der Endposition des VF8 Musters ist nicht im MeMotif Muster enthalten. Von diesem Standpunkt aus gesehen ist es daher letztendlich unverständlich, weshalb die Paarungen GG4VF8, GG4IV4 und IV4VF8 so wenig vertreten sind, gerade weil das GG4 Muster schon bei Gerstein et al. mit zahlreichen Funktionen in der Transmembran in Verbindung gebracht wurde, vor allem in der Rolle als Glycinzipper.

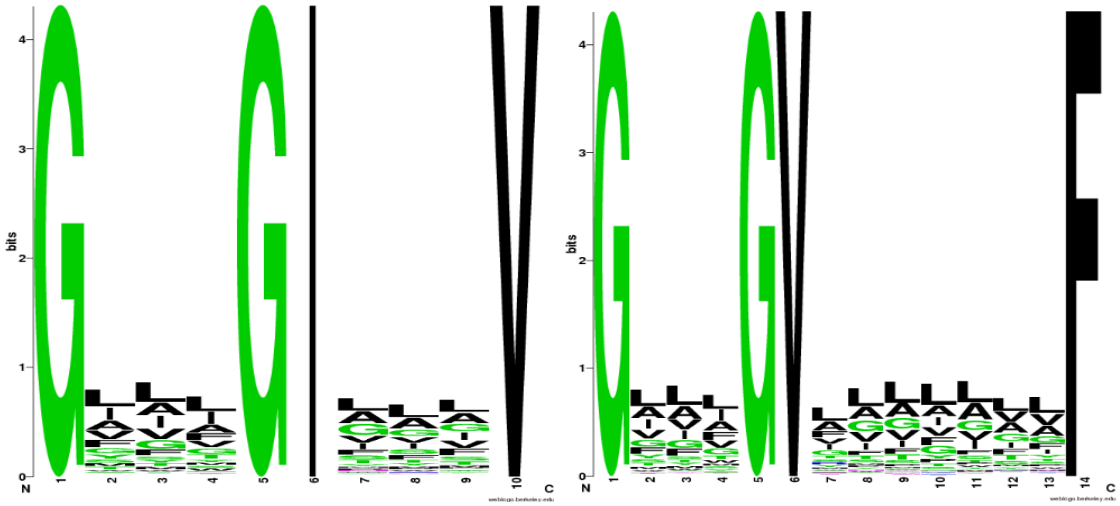

Abbildung 23 WebLogos für GG4IV4 und GG4VF8 im TM – Bereich

Abgebildet sind die WebLogos der Ausprägungen der Paarungen GG4IV4 und GG4VF8 aller 32 Proteinfamilien im transmembranen Bereich.

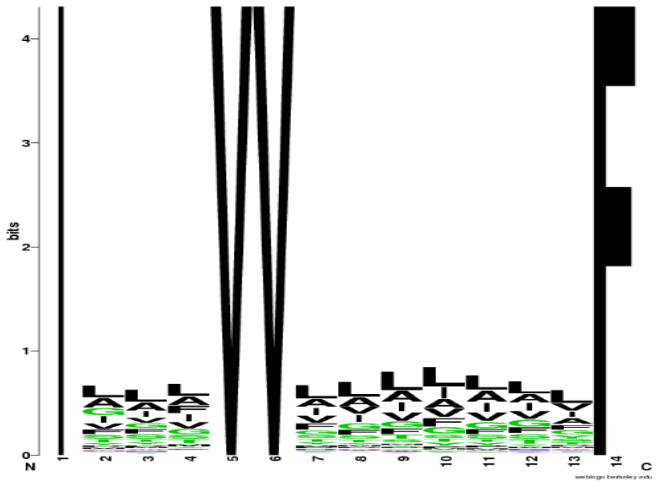

Abbildung 24 WebLogo für IV4VF8 im TM – Bereich

Abgebildet ist das WebLogo der Ausprägungen der Paarung IV4VF8 aller 32 Proteinfamilien im transmembranen Bereich.

Auch bei Betrachtung der WebLogos der Ausprägungen der drei Musterpaarungen (siehe Abbildung 23 und 24) ist zu erkennen, dass sie eigentlich absolut typisch für den transmembranen Bereich sind, da auf den Zwischenpositionen am häufigsten Leucin und Alanin zu finden sind, daneben Isoleucin und bei der Paarung IV4VF8 kommt auf der zweiten Position als dritthäufigstes Glycin vor, was wiederum eine unterstützende Funktion für die Proteinstruktur innehat. Die tatsächliche Funktion der drei Paarungen müsste also in weiterführenden Untersuchungen genauer untersucht werden.

Abschließend wurde noch einmal versucht, auch Paarungen, die nur in einzelnen Familien gehäuft auftreten, eine mögliche Funktion zuordnen zu können und so stellte sich heraus, dass dies für die Paarung GA4GL1 im transmembranen Bereich der Familie PF09767 der Fall sein könnte. Wie bereits vorher erwähnt, wurde GL1 in einem Loopmotif von MeMotif (GLYYGSY) gefunden, welches zum einen zwei Helices miteinander verbindet und zum anderen gleichzeitig Teil der Bindungstasche für die Hämgruppe ist, was dahingehend interessant ist, dass GA4 in einem Muster gefunden wurde, welches auch die Hämbindung involviert ist und zwar handelt es sich hierbei um das MeMotif Muster QI[LV]TG[IL][FLV][LM]AMHY. Zwar ist der Loop, in dem GL1 gefunden wurde nicht der TM – Helix zugehörig, jedoch kann man aufgrund dem Fakt, dass die Paarung GA4GL1 auch gehäuft auftritt, wenn man die einzelnen Bereiche bei der Mustersuche nicht differenziert, aussagen, dass diese Paarung, von der ein Muster im Transitionbereich und eins in einer Helix auftritt, möglicherweise im Zusammenspiel dafür sorgen, dass die Hämgruppe binden kann. Für weiterführende Untersuchungen wäre es in diesem Zusammenhang möglicherweise durchaus von Vorteil, wenn man die MeMotif Muster als Ausgangspunkt nimmt, um so genauere Aussagen über Funktionen von gemeinsam auftretenden Mustern treffen zu können. In diesem Zusammenhang wurde auch untersucht, ob das GLYYGSY Muster von MeMotif, welches in 909 Transmembranproteinen in der Swiss-Prot Datenbank gefunden wurde, in der Pfamfamilie PF09767 vorkommt, dies war jedoch nicht der Fall. Wäre der Fall jedoch eingetreten, hätte man das entsprechende Protein möglicherweise einer bereits beschriebenen Familie zuordnen können.

Abbildung 25 WebLogo für GA4GL1 in Familie PF09767

Abgebildet ist das WebLogo für die Ausprägungen des GA4GL1 Musters im TM – Bereich in der Pfamfamilie PF09767.

Das WebLogo in der Abbildung 25 zeigt, dass das GA4 Muster der Paarung an zweiter Position am häufigsten ein Asparagin enthält, was dahingehend interessant ist, da es sich hierbei um eine hydrophile Aminosäure handelt. In wie weit dieses Asparagin Einfluss auf Funktion oder Struktur nimmt, lässt sich jedoch nicht sagen, zudem ist die Beobachtung nur einer einzelnen Proteinfamilie nicht sehr aussagekräftig, wenn man Funktionen verallgemeinern möchte.

4.5 Fazit

Für den Transitionbereich wurden verschiedene Musterpaarungen wie AA3GL1, AA3AG4, GL1GL3, GL1GL2, AA3AL6 und AL6GL1 gefunden, die zum einen teilweise relativ oft in diesem vorkommen (AL6 Muster) und zum anderen womöglich eine bestimmte Funktion innehaben. Im Verbund könnten diese in Reentrantloops für die Stabilität des Proteins sorgen, da sie die mit am häufigsten vorkommenden Paarungen miteinander bilden. Ob dies jedoch tatsächlich der Fall ist, muss noch untersucht und belegt werden.

Im TM - Bereich fanden sich nur einzelne Musterpaarungen, denen eine Funktion zugeteilt werden konnte, diese beziehen sich vor allem auf Helix - Helix Interaktionen, was besonders wichtig für die Struktur des gesamten Proteins ist. Das Problem bestand darin, dass sich die Funktionen der Datenbankmuster zum größten Teil auf Reentrantloops bezogen oder auf Prozesse, die nichts unmittelbar mit dem Transmembranbereich und vor allem der Struktur zu schaffen haben.

Insgesamt gesehen ließ sich auch dem häufigstem Musterpaar, dem AA2AA3, keine Funktion zuordnen. Entweder ist darüber noch nix bekannt oder aber das AA2 Muster ist aufgrund seiner geringen Länge nicht wirklich in Transmembranbereichen konserviert, sondern nur so oft vertreten, da zwei von drei Positionen von Alanin besetzt werden, welches auch in der Natur neben Leucin die am häufigsten vorkommende Aminosäure ist, wodurch die Wahrscheinlichkeit, dass ein Alanin, egal in welchem Bereich eines Transmembranproteins, besonders häufig vorkommt.

5 Ausblick

Im Laufe dieser Arbeit konnte für einzelne Muster aus den Arbeiten von Gerstein et al. gezeigt werden, dass sie eine bestimmte Funktion innezuhaben scheinen, auch oder vor allem, wenn sie zusammen mit anderen Mustern innerhalb einer Proteinsequenz vorkommen. Weiterführende Untersuchungen und Arbeiten sollten sich vor allem mit den Mustern beschäftigen, die aus mindestens 4 Aminosäuren bestehen, da man erst diese besser zu bestimmten Bereichen der Transmembranproteine zuordnen kann. Zudem erscheint es sehr wichtig, sich die Zwischenpositionen ausgewählter Paarungen noch genauer anzuschauen, um zu sehen, welche Aminosäuren denn überhaupt maßgeblich über eine bestimmte Funktion des Musters entscheiden.

Des Weiteren erscheint es nur logisch, die Co - Co - Occurence auf drei, vier oder mehr Muster auszubauen, um möglicherweise feststellen zu können, wie mehrere Muster, die sich in verschiedenen Abschnitten befinden, miteinander die Struktur eines Proteins beeinflussen und für deren Stabilität sorgen.

6 Literaturverzeichnis

Literatur/Internet

[1] Nöldner, Anne: Korrelation von Sequenzmotiven mit Prosite-Motiven in Membranproteinen. - 2010. – 82 S. Mittweida, Hochschule Mittweida, Fakultät Mathematik, Naturwissenschaften, Informatik, Bachelorarbeit, 2010

[2] Wellcome Trust Sanger Institut <contact@sanger.ac.uk>: Pfam, URL: <http://pfam.sanger.ac.uk/search/keyword?query=transmembran+proteins&submit=Submit>

[3] Pfam Home page, Wellcome Trust Sanger Institut <contact@sanger.ac.uk>: Pfam, URL: <http://pfam.sanger.ac.uk/>

[4] MeMotif:_Motifs_in_alphahelical_membrane_proteins – MeMotifWiki, Annalisa Marsico, Bioinformatics group Biotechnology Center TU Dresden, URL: <http://projects.biotec.tu-dresden.de/memotif/en/MeMotif:_Motifs_in_alphahelical_membrane_proteins>

[5] ELM, search the ELM resource, URL: <http://elm.eu.org/search/>

[6] ExPASy Prosite, prosite@expasy.org, URL: <http://prosite.expasy.org/>

[7] CD-HIT Suite, Weizhong Li, URL: <http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit>

[8] BLASTclust, Dept. of Protein Evolution at the Max Planck Institute for Developmental Biology, URL: <http://toolkit.tuebingen.mpg.de/blastclust>

[9] TMHMM Server, v. 2.0, cbs@cbs.dtu.dk, URL: <http://www.cbs.dtu.dk/services/TMHMM/>

[10] Regexp::Genex – search.cpan.org, Brad Bowman, URL: <http://search.cpan.org/~bowmanbs/Regexp-Genex-0.07/lib/Regexp/Genex.pm>

[11] Liu, Yang; Engelmann, Donald M.; Gerstein, Mark: Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biology 2002, 3(10):research0054.1–0054.12 2002

[12] Adamian, L. et al.,Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins,Journal of molecular biology,Vol.311:891-907,2001

Abbildungen

Abbildung 1 Scheffel-Gymnasium: BIO-LK - Proteine Enzyme,Werner Lingg – lingg@scheffel-gymnasium.de,URL: <http://www.scheffel.og.bw.schule.de/faecher/science/biologie/proteine_enzyme/1protein/amino.gif>

Abbildung 2 Wikipedia,URL: <http://de.wikipedia.org/wiki/Datei:Amino_Acids_Venn_Diagram_%28de%29.svg>

Abbildung 3 TU Leipzig, URL:<http://www.bioinformatik.unileipzig.de/Leere/PRAKTIKUM/Protokolle/SS04/1/Bilder/protein_einf.jpg>

Abbildung 4 Chris Brauer, URL: <http://chrisbrauer.wikidot.com/mb:pruefungsfragen>

Abbildung 5 Pfam Homepage, <contact@sanger.ac.uk>:, URL: <http://pfam.sanger.ac.uk/>

Abbildung 6 Motifs_in_alphahelical_membrane_proteins – MeMotifWiki, Annalisa Marsico, Biotechnology Center TU Dresden, URL:<http://projects.biotec.tudresden.de/memotif/en/MeMotif:_Motifs_in_alphahelical_membrane_proteins>

Abbildung 7 ELM, URL: <http://elm.eu.org/elms/browse_elms.html>

Abbildung 8 ExPASy Prosite, prosite@expasy.org, URL: <http://prosite.expasy.org/cgi-bin/prosite/prosite-search-ac?PDOC00284>

Abbildung 9 CD-HIT Suite, Weizhong Li, URL: <http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi>

Abbildung 11 G. E. Crooks et al.,<logo@compbio.berkeley.edu>,

Tabellen

Tabelle 1 Anne Nöldner: Korrelation von Sequenzmotiven mit Prosite-Motiven in Membranproteinen, Mittweida, Bachelorarbeit, 2010

Anhang

(1) Tabelle aller 50 verwendeten Muster und deren Anzahl in den TM – und nTM – Bereichen

Muster	nTM	TM
AA2	4935	6530
AA3	4998	6644
AA7	4458	5933
AG3	3926	4360
AG5	3552	4348
AG7	3504	4389
AI	2101	4456
AL6	4345	8754
AS4	3224	2428
FG	1680	2517
GA4	3721	4518
GA7	3501	4227
GG4	3105	2568
GG7	2801	2966
GL	3359	6832
GL3	3620	6628
GN4	1223	324
GS4	2705	1420
GS5	2564	1481
GV	2383	3597
GV2	2782	3675
GY8	1117	1248
IA	2246	5053
IG	1845	3778
IL4	2535	6895
IV4	1673	4001
LF8	1896	5294
LF9	1907	5327
LF10	2039	5062
LG5	3564	6017
LP	3273	2436
LS	3928	4314
LY6	1576	2771
PF	1042	851
PF2	1120	675
PG5	1954	819
PG6	2205	808
PG9	2056	878
PG10	2058	778
PL	2893	2370
PL2	2987	2387
(SA3)	3310	2388
(SA6)	3162	2392
SG4	2771	1449
SL	3669	4180
SL2	3649	4035
VF8	1305	2929
VG6	2618	3417
VL4	3349	7934

(2) Übersicht über alle Muster, die in Datenbankmustern konserviert vorliegend gefunden worden.

Muster	gefunden in	Funktion
AA2	EGYATA	part of the active center for primase function
AA3	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
AG4	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	protein structural stability and dimerization
	A-[GL]-[LM]-[AS]-A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I	reentrant loop in chloride channels
AL6	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	protein structural stability and dimerization
AG7	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	T-A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G	part of the binding pocket of heme b and the close inhibitor stigmatellin
	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	A-[LMF]-x-[GAT]-T-[LIVMF]-x-G-x-[LIVMF]-x(7)-P	This protein probably forms a transmembrane proton channel
FG1	F-F-G-V-x(2,3)-F	Membrane-water interface motif
	F-F-G-V-x-[AGT]-[FIM]	interactions with cofactors, lipid-exposure motif
GL1	G-L-Y-x-G	iron ion binding, electron transport
	G-L-Y-Y-G-S-Y	Structured-loop motif connecting two helices
	N-P-[AV]-P-[LF]-G-L-x-[GSA]-F	could be involved in transport
GL2	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	V-G-x-L-[ACTV]-[AGT]-[IL]-[AGV]	involved in helix-helix packing
	V-x(0,2)-G-x(3)-G-x(1,2)-L	-
GV2	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
GL3	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	G-V-x(1,2)-L-[ILV]-[GL]-[TV]-[LV]-[LM]	putative helix-helix interaction
	G-[CGS]-[AG]-L-[FL]-x-A-x-H-G-[AS]-x-[IV]	nvolved in the binding of the mononuclase of M chain
GA4	L-G-[ILM]-G-x-H-x(1,3)-A-x(0,2)-F	conserved site for binding of cardiolipin
	Q-I-[LV]-T-G-[IL]-[FLV]-[LM]-A-M-H-Y	involved in heme binding
	EGYATA	part of the active center for primase function
GG4	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	G-L-Y-x-G	iron ion binding, electron transport
	G-L-Y-Y-G-S-Y	Structured-loop motif connecting two helices
	T-A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G	part of the binding pocket of heme b and the close inhibitor stigmatellin
	A-F-[LMV]-G-Y-x(0,2)-V-x(1,3)-G-Q-M-S	electron transport chain
	G-G-x-[LIVM]-G-[LIVM]-x-[IV]-x-W-x-C-[DN]-L-D-x(5)-C-x-P-x-Y-x-F	P2X purinoreceptors are cation-selective ion channels
	C-D-G-P-[GE]-R-G(2)-T-C	The 2 C's bind the iron-sulfur center
	[LIVMF]-x-G-[LIVMFA]-{V}-x-G-{KP}-x(7)-[LIFY]-x(2)-[EQ]-x(6)-[RK]	contains many positive charges and serine residues that could be phosphorylated
	[GD]-x(0,10)-[FYWA]-x(0,1)-G-[LIVM]-x(0,2)-[LIVMFYD]-x(0,7)-	share the presence of a conserved, glycine-rich domain
	G-[KN]-[NHW]-x(0,1)-G-[STARCV]-x(0,2)-[GD]-x(0,2)-[LY]-[FC]	
	LIXXGVXXGVXXT	dimerization motif for transmembrane alpha-helices
	[GAS]XXXGXXXG	distinct preference for right-handed packing
GG7	[LM]-[LF]-T-x-R-[SA]-x(3)-[RK]-x(3)-G-x(3)-F-P-G(2)	involved in CoA recognition
	E-x-G-G-P-x(2)-[GA]-x-G-C-[AG]-G	nucleotide-binding proteins [C binds the iron-sulfur center]
GN4	G-[AGT]-x-[FIM]-N-P-A-x-[ST]-[FI]-[AG]	selectivity filter motif in water-glycerol transporters
GS5	G-L-Y-Y-G-S-Y	Structured-loop motif connecting two helices
	GXXRXSG	arginine finger
GV1	G-V-x(1,2)-L-[ILV]-[GL]-[TV]-[LV]-[LM]	putative helix-helix interaction
	F-F-G-V-x(2,3)-F	Membrane-water interface motif
	LIXXGVXXGVXXT	dimerization motif for transmembrane alpha-helices
IA1	F-E-D-[LV]-I-A-[DE]-[PA]	Caveolins may act as scaffolding proteins within caveolar membrane
IV4	LIXXGVXXGVXXT	dimerization motif for transmembrane alpha-helices
LG5	L-x-[AIV]-x-[HW]-G-A-T	modulator of transmembrane protein function
	F-x(3,4)-L-x(1,2)-L-[IL]-G	structural and involved in interactions with the bilayer lipids
LF8	L-G-[ILM]-G-x-H-x(1,3)-A-x(0,2)-F	conserved site for binding of cardiolipin
	L-[DP]-[FLV]-x-[AST]-K-V-G-F-[GS]	part of the retinal binding pocket of bacteriorhodopsin
	D-x-L-G-D-V-V-C-G-G-F-[AGSP]-x-P	nucleotide-binding proteins [C binds the iron-sulfur center]
PG5	P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]	Shorter version of the P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]-[AGST]-[FILV].
	P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]-[AGST]-[FILV]	iteraction with different membrane lipids and thecytochrome subunit
PG6	L-W-x-[AGI]-Y-P-[FIV]-[ILV]-W-[IL]-[ILV]-G	Proline-kink motif
PG10	ISXP(X)2,4-FXHXXHVG	confers binding to the Cdc42 and/or Rac GTPases
PL2	F-F-[DQ]-P-[AS]-G-G-G-D-P-[IV]-L-Y	Glycines confers flexibility to the whole structural motif
PL1	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I-x(0,1)-L	Reentrant loop in chloride channels
	L-x-A-A-x(2,3)-A-P-L-[AT]-G-I	Reentrant loop in chloride channels
	P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]	Shorter version of the P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]-[AGST]-[FILV].
	P-L-x-[DEK]-G-G-x(0,1)-W-x(2)-[AI]-[AGST]-[FILV]	iteraction with different membrane lipids and thecytochrome subunit
	A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G	protein structural stability and dimerization
	[DET]E[RK].PL[LI]	interacting with the AP3 adaptor complex
	A-[GL]-[LM]-[AS]-A-[AS]-[FIV]-[NR]-A-P-L-[AT]-G-I	reentrant loop in chloride channels
SG4	IYTSGTTGXPK	interact with ATP

7 Selbständigkeitserklärung

Hiermit erkläre ich, dass die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel von mir angefertigt wurde.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.
Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, den 22. August 2012

Martin Garbe